Technological Advances in Corpus Sampling Methodology

Stephen Michael Wattam, BSc (Hons), MRes

Mathematics and Statistics

School of Computing and Communications

Lancaster University

Submitted for the degree of

Doctor of Philosophy

August 2015

Abstract

Current efforts in corpus linguistics and natural language processing make heavy use of corpora—large language samples that are intended to describe a wide population of language users.

The first modern corpora were manually constructed, transcribed from published texts and other non-digital sources into a machine-readable format. In part due to their hard-won nature, larger corpora have become widely shared and re-used: this has significant benefits for the scientific community, yet also leads to a stagnation of sampling methods and designs.

The rise of Web as Corpus (WaC), and the use of computers to author documents, has provided us with the tools needed to build corpora automatically, or with little supervision. This offers an avenue to re-examine and, in places, exceed the limitations of conventional corpus sampling methods. Even so, many corpora are compared against conventional ones due to their status as a de-facto gold standard of representativeness. Such practices place undue trust in aging sample designs and the expert opinion therein.

In this thesis I argue for the development of new sampling procedures guided less by concepts of linguistic balance and more by statistical sampling theory. This is done by presenting three different areas of potential study, along with exploratory results and publicly-available tools and methods that allow for further investigation.

The first of these is an examination of temporal bias in sampling online. I present a preliminary investigation demonstrating the prevalence of such effects, before describing a tool designed to reveal linguistic change over time at a resolution not possible with current software.

Secondly, the sample design of larger general-purpose corpora is inverted in order to relate it to an individual's experience of language. This takes the form of a census sample of language for a single subject, taken using semi-automated methods, and illustrates how poorly suited some aspects of general-purpose corpora are for questions about individual language use.

Finally, a method is presented and evaluated that is able to describe arbitrary sample designs in quantitiative terms, and use this description to automatically construct, augment, or repair corpora using the web. This method uses bootstrapping to apply current sampling theory to linguistic research questions, in order to better align the scientific notion of representativeness with the process of retrieving data.

Acknowledgements

This thesis could not have been completed without the support of many people. Through the considerable time it took, many academics and friends have helped answer my questions and allay (or confirm!) my fears. My thanks go out to all members of UCREL, who taught me the ways of the corpus.

Foremost amongst these must be my supervisors, Paul Rayson and Damon Berridge. Their direction and encouragement was often the only thing keeping me in the office, and it is hard to overstate their influence.

Secondly, my friends, particularly Carl Ellis, Matthew Edwards, John Vidler and John Hardy. Our (sometimes not so) implicit competition has been endlessly motivating.

Finally, I wish to thank my parents. It is doubtless my mother's lesson to question the world that has lead me into research, and my father's work ethic (however diluted during inheritance) that has carried me this far.

Thank you all.

Declaration

The material presented in this thesis is the result of original research by the named author, Stephen Wattam, under the supervision of Dr. Paul Rayson & Prof. Damon Berridge, carried out in collaboration between Maths and Stats and the School of Computing and Communications, Lancaster University.

This work was funded by the Economic and Social Research Council. Parts of this thesis are based upon prior publications by the author, listed below—all other sources, if quoted, are attributed accordingly in the body of the text.

- Stephen Wattam, Paul Rayson & Damon Berridge. Document Attrition in Web Corpora: an Exploration. In *Corpus Linguistics*, 2011.
- Stephen Wattam, Paul Rayson & Damon Berridge. LWAC: Longitudinal Web-as-Corpus Sampling. In *8th Web as Corpus Workshop (WAC-8)*, 2013.
- Stephen Wattam, Paul Rayson & Damon Berridge. Using Life-Logging to Re-Imagine Representativeness in Corpus Design. In *Corpus Linguistics*, 2013

Mr Stephen Wattam .....

Dr Paul Rayson .....

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Kuhn's view of scientific revolutions[105] holds that only empirical evidence, well applied, may provide the theoretical framework required to progress a field; as the intricacy of theoretical systems evolves, this necessarily requires more accurate data. Like many of the younger social sciences, linguistics faces a difficult dilemma: it must have high-quality empirical data in order to further develop widely-agreed-upon theories, yet its ability to establish the quality of samples is directly dependent upon said theories.

Current corpus resources are often fashioned around a small set of exemplar corpora—these have become de-facto standards, and their sample designs are heavily re-used. Whilst this approach ensures compatibility and interoperability, it also leaves many sample design decisions unspoken or assumed.

Many of these examplar corpora were built in the days before widespread use of technology such as the internet, before digital document creation became popular. As a result they are structured around designs intended to minimise practical issues surrounding retrieval, coding, and transcription that may not apply in a modern context. These challenges had the effect of prying corpus designs from their ideal statistical principles and into more expert-designed forms[7].

This thesis presents three explorations of this gap: identifying methods and techniques for sampling in unconventional ways, and assessing their worth within the context of existing general-purpose samples for linguistic research. For each, a generalisable method and supporting tools are produced, in order to make these efforts accessible to a wider linguistic audience.

At its core is the notion of representativeness, the central argument being that the key to improving the validity of corpus-based scientific research (in linguistics or NLP), is the improvement of the sampling method.

Research Questions

RQ1 How can existing sampling theory from other disciplines be used to extend current general-purpose corpus construction techniques?

RQ2 How can WaC methods be applied to investigate and explore issues of representativeness?

RQ3 How can WaC methods be used to mitigate existing challenges in general-purpose corpus design?

Background — Chapter 2

One significant issue, which remains a topic for debate, is what goals linguistic samples should have. As the focus in this thesis is on larger, general-purpose, corpora, this requires some understanding of common use, and existing practice.

This chapter begins by reviewing the history of general-purpose corpus building, with a view to constructing a working definition of a corpus using existing terminology from corpus linguistics. These designs are then compared to a number of relevant 'ideal' sample designs informed by statistical sampling theory, and their similarities and limitations discussed. The closing portion of the review forms a brief introduction to new technologies available, particularly the Web-as-Corpus paradigm.

Document Attrition — Chapter 3

As corpora involving web data become more prevalent, the need to understand the temporal properties of web publishing become increasingly important to the external validity of any scientific claims made using them. Many others have already observed these changes, though with a focus on technical, rather than linguistic properties.

This chapter presents a motivating investigation into sources of link rot (the tendency for resources online to become unavailable over time) within corpora distributed as lists of URLs. It finds that many corpora suffer significant loss when retrieved using naive methods, across a timescale likely to affect users of current corpora (some widely-used corpora are now decades old).

This motivates long-term and in-depth analyses, neither of which I am well placed to cover in this thesis: however, a comprehensive tool was developed in order to perform longitudinal sampling of web data. This tool offers a mechanism for investigating properties as yet unaddressed by the literature, such as changes through time that do not affect availability (such as the editing of articles according to political whim), and network-level properties (such as the response time and location of hosting providers).

Contributions:

- A minor addition to existing literature on link rot, particularly in linguistic corpora.
- A publicly-accessible tool for longitudinal analysis of web resources, at a resolution and scale not possible with current software.

Personal Corpora — Chapter 4

The issue of representativeness is examined here with respect to an unconventional research question: how representative is a general-purpose corpus of just one speaker's experience of a language?

This question is valuable primarily because it inverts the sample design commonly used for larger corpora, which cover many people in many contexts but with a low resolution for each individual. Inverting the standard design allows for detection of a number of difficult-to-determine properties, such as what proportions of each genre are used at given times of day, or in different contexts.

At the same time, many of the challenges faced when executing such a sample design are novel, or have only been seen outside corpus sampling. Here, too, technological benefits such as the ubiquity of smartphones and speed of logging software yield results that would have been impossible just a few years ago.

This section ends with a discussion of the potential uses of such a sample, both of itself (for use in special-purpose corpus construction) and at a wider scale to inform the sampling policy of future efforts (to identify missing components from typical designs).

Contributions:

- A case-study relating general-purpose corpora to one person's idiolect; the practical and theoretical findings therefrom.
- A semi-automated method for producing personal corpora, including supporting software.

Profile-Guided Corpora — Chapters 5, 6

A central theme of this thesis is the research-question-oriented nature of any sample. Without a clear and objective research question, any claim to external validity is subject to the judgement of the reader.

These two chapters detail the design and implementation of a method that is designed to unambiguously specify the important variables in a research design, and then retrieve corpora based upon these designs. This serves to solve a number of significant practical issues with conventional corpus design, and takes sampling from the web in a new direction: away from a corpus of the web itself, and towards a general-purpose source of documents for all purposes.

This is accomplished by the encoding of procedures based on expert opinion into pluggable modules, that may be used to construct a corpus' sample design according to the whim of the user. These designs may then be distributed free of any legal restriction and re-used at will. The automated nature of the corpus retrieval process also serves to vastly speed up and simplify corpus construction.

Because there is no reliance on a gold-standard corpus, evaluation for this method is particularly difficult. This thesis relies on two mechanisms: a white-box examination of each of

the stages of the corpus building process is presented, and its rationality confirmed. Secondly, a black-box examination against the source corpus is then performed, in order to identify any subjective unintended bias in the result. Though the results of this evaluation are difficult to generalise to all corpus designs, these indicate a practical level of success for the common design used.

Contributions:

- A method for quantitatively producing, reweighting, and describing corpora using a set of metadata distributions using bootstrapping.
- Software tools to summarise corpora into freely-reusable profiles.
- Software tools to construct corpora based upon said profiles, using fully-automated and semi-automated methods.
- A mechanism to mitigate existing ethical and legal issues surrounding corpus distribution.

Conclusions Chapter 7

This thesis concludes with a review of the methods presented, and an examination of the contexts in which they may be used. The core themes of the thesis are reviewed, before a discussion of major items of further work.

Chapter 2

Background

Large samples of text, known as corpora, are important for many users. The groups I will be focusing on for the purposes of this thesis are particularly interested in scientific samples, that is, in generalising findings from the corpus to a wider population. This includes much work in Natural Language Processing (NLP) and corpus linguistics, and some of what is commonly termed 'data science'[1].

The problem of sampling linguistic data is a well-researched one in the field of corpus linguistics, and a review of linguistics' approach to corpus construction will form the bulk of this chapter. This portion of the review begins with a historical overview of approaches to corpus construction, before focusing on the problem of defining a corpus in sufficiently concrete terms to suit the rest of the work presented in the thesis. It then focuses on potential threats to validity exposed by these methods, both from the point of view of corpus construction and use.

Not to be neglected, however, is the perspective offered by more formal survey sampling, especially as its use across the social sciences can offer insights into solutions for practical problems. In Section 2.2 a number of appropriate sample designs are identified and evaluated with respect to corpus sampling.

Finally, in Section 2.3 the first few steps towards improving sampling methodology within corpus linguistics are reviewed, with a focus on web-based corpus construction methods.

2.1 Corpora and Corpus Linguistics

The field of linguistics is one concerned with description and formalisation of a particularly ethereal social concept. The paucity of philosophical agreement upon the nature of language has led to many different approaches being taken through the years, many of which have accomplished great things in advancing our capacity to reason about, and derive conclusions from, language.

The most obvious method for inquiring about the nature of language is to sample real-world use. This process is followed in many other sciences concerned with social phenomena, and

[1] Though in most 'data science' contexts, sampling is performed as an intrinsic part of the task

offers a tried-and-tested methodology for inferring results. It is perhaps unsurprising, then, that this method has been used to a varying degree by many linguists throughout history.

In their 2001 book[127], McEnery and Wilson characterise the pre-Chomskian linguistic inquiry as primarily taking this form:

> *The dominant methodological approach to linguistics immediately prior to Chomsky was based upon observed language use.*

They point to a number of studies using systematic analysis of language samples to draw their conclusions conclusions, ranging from the late 19th century through to the "modern era" of corpus analysis[90, 141, 170, 42, 188, 182]. They also decribe Chomsky's influences on the field, which prompted to lead it to shun corpus techniques in the 1950's and turn towards explanatory, rationalist theories of language.

These rationalist theories were often verified using experimentation or elicitation, in an effort to gather data that is detailed and reliable. It's arguable just how valid and philosophically defensible these methods are, especially given the nature of language as a part-mental, part-tangible concept.

In a pre-digital world, collection and analysis of large-enough-to-be-useful samples of language was extremely difficult, making this rationalist approach a viable alternative. Small samples are fundamentally unable to reveal some of the details examined by the structuralists, and construction and analysis of sufficiently large corpora in a non-digital era would prove a practical hard limit on the power of corpus studies. To frame the focus on rationalist inquiry as an alternative to the empirical is a disservice to both: empiricism had supported the rationalist theories of the 50's, and would itself go on to be supported in turn as it once again rose to prominence.

The renaissance of corpus linguistics may be attributed primarily to the availability of programmable computing machinery. This offered a solution to the problems of scale encountered during earlier corpus analyses, making empirical data once more viable for detailed inspection of language.

This revival, often termed the 'modern era' of corpus linguistics, has yielded what we would commonly call a corpus today: a large, machine-readable, annotated collection of texts sampled in order to represent some population of real-world language use.

Corpus studies are now widely relied upon across linguistics, and are often the method chosen to test theories derived from structuralist approaches. This may be seen as a validation of their methods, as the two complementary philosophies of scientific inference are once more able to use one another without methodological suspicion.

For the purposes of this section I will be focusing on the design of 'modern' corpora with respect to their use in validating linguistic theories, and for training automated NLP and Information Retrieval (IR) systems. For this reason I will be covering mainly general purpose corpora: those that purport to represent such a large population as to cover a whole language for most research questions. This type of corpus is built so as to be useful to many research questions and researchers, in part to dilute practical issues surrounding sampling. This is in contrast to

special-purpose corpus, which are designed to represent a restricted context. Special-purpose corpora may be selected according to demographic or linguistic properties, and are typically much smaller. Because of this they are often built for a given study, or by re-sampling a general-purpose corpus.

2.1.1 A Brief History of Modern Corpora

The Brown Corpus of Standard American English[52]is widely regarded as the first of the modern age of corpora. Built in the 60's, Brown's corpus was the first electronic-format general purpose corpus and was roughly one megaword in size. It contains 500 samples, each comprising roughly 2,000 words, that were taken to represent a cross-section of works published in the United States in 1961. The proportions and sizes of samples were selected in order to trade off pragmatic concerns with the possible kinds of analysis that could be performed at the time.

The 'Standard' in its name referred to Kucera and Francis' intent that the corpus represents their judgement of 'standard' English use. Brown became a *de facto* standard for American English, and the design was carried forward into many other corpora, mostly regional versions designed to be comparable to Brown[78, 87, 77, 162, 35, 17, 128]. In order to maximise the value of comparisons within studies, other general purpose corpora chose to mirror Brown's sampling policies.

Categories	Texts in each category	American corpus	British corpus
A	Press: reportage	44	44
B	Press: editorial	27	27
C	Press: reviews	17	17
D	Religion	17	17
E	Skills, trades, and hobbies	36	38
F	Popular lore	48	44
G	Belles lettres, biography, essays	75	77
H	Miscellaneous	30	30
J	Learned and scientific writings	80	80
K	General fiction	29	29
L	Mystery and detective fiction	24	24
M	Science fiction	6	6
N	Adventure and western fiction	29	29
P	Romance and love story	29	29
R	Humour	9	9
	Total	500	500

Table 2.1: The basic composition of the British and American corpora

The Lancaster-Oslo-Bergen (LOB) corpus[87] was built as a British counterpart to Brown. It uses the same stratification and sampling strategy (with one or two more texts in certain categories) and thus comprises roughly a megaword of British English, as published in 1961. The manual notes:

> *The matching between the two corpora is in terms of the general categories only. There is obviously no one-to-one correspondence between samples, although the general arrangement of subcategories has been followed wherever possible.*

Table 2.1 shows the proportions of texts in each genre, relative to Brown (the 'American corpus'), as reproduced from the corpus manual.

The London-Lund Corpus[63] was released in 1990, and contains transcribed spoken text, annotated with a number of different markers to indicate intonation, timing and other extra-textual information. The corpus consists of 100 texts, each of 5,000 words, totalling 500,000 running words of spoken British English. The annotation scheme used in LLC is much more in-depth than that in many written corpora, reflecting the different nature of research questions for speech.

Collins' requirement for a corpus upon which to base their dictionaries spawned the COBUILD and its 'representative' subset, the Bank of English[85, 164]. COBUILD uses a slightly different approach to corpus building: that of the monitor corpus. Monitor corpora are continually added to, using a fixed sampling policy but an ongoing sampling process. At the time of writing, the BoE is 650 million words in size (the whole COBUILD family used by Collins is 4.5 billion)[109].

The approach taken by the BoE opens many possibilities for analysis of language over time (something also covered by diachronic/historical corpora using more conventional sampling). Even so, such comparisons are complicated by the irregular additions (c.f. diachronic corpora, which contain complete samples for each time), and this is the only major corpus built in this fashion prior to automated web retrieval.

The de-facto standard of the day is currently the British National Corpus, which comprises 100 million words of British English[114]. The BNC's design was influenced heavily by discussions on corpus building that centred around creating a standard, reliable approach to taxonomy, sampling and annotation that occurred around the early nineties.

The BNC aims at being a synchronic 'snapshot' of British English in the early 1990s. It consists of samples of text up to 45,000 words each, and is deliberately general-purpose, containing a wide range of genres as well as a sizable spoken portion. It was released in 1994, but has since been re-coded and augmented, particularly notably by Lee[111], who constructed a significantly more detailed (and principled) taxonomy for its texts in 2003.

Lee's genre taxonomy is outlined in Appendix C. Lee's decisions to code genres using this system were partially based on fostering compatibility with the ICE-GB[62] and LOB[87] corpora.

The prominence and 'whole population' coverage of the BNC's sampling frame spawned many compatible corpora, though these often show more variation in sample design than the Brown clones. Xiao, in his survey of influential corpora, notes the existence of national reference corpora for American, Polish, Czech, Hungarian, Russian, Hellenic, German, Slovak, Chinese, Croation, Irish, Norwegian, Kurdish, and many more[186]. An updated version, BNC2014[140], is currently being constructed.

The rise of electronic communications has led to a reduction in the effort required to gather corpus data. This has resulted in a great increase in the number of special-purpose corpora built for specific studies[157, 16, 121]. These corpora are more focused in that their construction methods restrict them to electronic data, yet their large sample size may make them suitable for the study of smaller-scale features in a more general context.

This thesis is focused on sampling using automated, technical mechanisms to overcome some of the challenges facing conventional methods, and as such focuses on web-based methods (known as 'Web as Corpus'). WaC is concerned with sampling the web itself, as well as constructing samples that are representative of other data, yet are retrieved primarily from the web. This approach, and those using similar methods, are covered below in Section 2.3.

2.1.2 What Makes a Corpus?

The use of general purpose corpora as large monoliths, reused in many studies and systems, has led to much debate over the nature of a corpus. This, as we shall see throughout the thesis, is a question unworthy of simple answers—each purpose will exert certain demands upon corpus design criteria, and any widely-used corpus is likely to be a compromise around these.

This section exists primarily to define the term 'corpus' as used here: it is unlikely that the definition derived below is universal, however, it is designed to reflect most use-cases, across corpus linguistics and NLP.

In order to establish the important traits of corpora, it is wise to have an understanding of the motivation behind their existence. Corpus methods are generally contrasted against two other methods of linguistic investigation: direct elicitation from a language speaker, and directed research into a linguistic feature under controlled conditions. Both of these pose significant scientific challenges—both are reliant on at least one linguist's intuitive view of language (one that could hardly be said to be representative of most language users), and both require the acquisition of data without its usual context, something that is especially difficult given the varied and context-dependent complexity of language.

Corpora provide limited solutions to both of these issues. In the former case, they provide an objective record of linguistic data that is free from all but the initial builders' influences (which, in the ideal case, may be documented and provided along with the data itself). In the latter, they are as portable as any large volume of text, and may be annotated with context sufficient for a given linguistic task.

At its most basic level a corpus is a sample of text. Given research questions surrounding a body of text, it is perhaps necessary only to stipulate that a corpus must be representative of a known population [127, p. 22]:

> *. . . a body of text which is carefully sampled to be maximally representative of a language or language variety.*

Further to this, the modern definition of a corpus has undergone a series of significant refinements thanks largely to the increase in both the ubiquity and power of computing machinery. General-purpose corpora are, with very few exceptions, machine-readable (with an increasing number documenting texts of electronic origin), multi-modal (covering a wide variety of methods of communication and their linguistic features), and annotated with linguistic data.

Many authors go further by stating that a corpus should be annotated with information useful to linguistic inquiry, built for a specific purpose or methodology, available for use in other studies,

finite in size or stratified to provide multiple possible analysis methods with internally valid data[19, 28, 113]. Whilst I do not consider many of these to be requirements for a scientifically useful corpus, they may contribute greatly to corpus utility due to their alignment with common methodologies and uses. There is undoubtedly a case for this—the utility of a corpus is often limited by its format, however, the extent to which this applies varies wildly by purpose makes it unreasonable to include many resources missing some of the above under the term 'corpus'.

This review, then, shall start with the most basic of definitions, that of 'a body of text sharing some property that may be interrogated for some linguistic information'. This takes into account the claims of generalisability that cannot be made for more haphazard collections of texts (often called *libraries* or *archives* with no relevant common features) which are not demarcated by the boundaries of some notional property.

Representativeness

Representativeness is, in effect, the goal of any sample. It is the property that I chose to use as the loose starting condition above, and it is to be maximised by selection of those properties covered below. By virtue of this holism, it is also dependent upon enough factors to be poorly defined in the literature.

The concept of representativeness is based strongly on philosophies of inference, and epistemology in general. These are highly dependent on correlation in variables external to those measured by a sample, and any rigorous definition will involve the properties we wish to generalise, the purpose of such generalisation, and the population we wish to generalise about. Users of reference corpora (those general-purpose corpora designed to represent a whole language) may find, for example, that their claims to validity are different to previous studies, simply due to the nature of their research question.

Much has been written on the subject of representativeness[23, 180, 179, 113, 145], both in linguistics and in other fields that are dependent upon complex sources of data. As with psychology or sociology, the opacity of the mechanisms that generate language is such that significant philosophical disagreement as to the underlying nature of the data occurs. This disagreement has, in many ways, limited efforts to formalise and reason about representativeness in corpus design: taken quantitatively, one person's adequate corpus is another's woefully biased one.

The LOB manual hints at the fluid nature of 'representativeness' in corpus linguistics, and the degree to which corpus design is expert-guided[87]:

> *The true "representativeness" of the present corpus arises from the deliberate attempt to include relevant categories and subcategories of texts rather than from blind statistical choice. Random sampling simply ensures that, within the stated guidelines, the selection of individual texts is free of the conscious or unconscious influence of personal taste or preference.*

I consider this a false dichotomy: a high quality random sample is essentially the gold standard to which expert designs should be aspiring, and both are aiming for the same notional

goal. The key benefit to random sampling is, as observed above, the immunity from 'unconscious' bias[12].

This ambiguity could be seen as an argument against the concept of a general purpose corpus, however, current progress indicates that this would be hasty: whilst special purpose corpora are often burdened by fewer procedural and pragmatic difficulties, they are necessarily limited in scope. It is still rational and useful to identify speakers of a language as a homogeneous group at some level, especially for smaller linguistic features. On the other hand, only a sample that properly encompasses a large amount of variation may be used to describe many effects of interest.

With this in mind, many of the studies into representativeness have focused on examining existing corpora, with a view to testing their internal variation against some known re-sample. This approach has been taken most famously by Biber[22], who performed a series of studies in which he compared the content of 100-word samples from the LLC and LOB corpora. Each sample was paired with another from the same text (LOB samples points are 2,000 words each, and LLC's are 5,000). Biber went on to extract some small-scale linguistic features from each sample, before examining the difference in frequency between each.

Biber concluded that existing corpora were sufficient for examination of smaller linguistic features, however, in the process he saw fit to reject the notion of representativeness used here (and, notably, everywhere else) [22, p. 247]:

> *Language corpora require a different notion of representativeness, making proportional sampling inappropriate in this case. A proportional language corpus would have to be demographically organized... because we have no a priori way to determine the relative proportions of different registers in a language.*

Biber's argument is that we should not aim for any degree of proportional representation of linguistic features, for this would produce a corpus that is mostly one kind of text (due to a conjectured Zipfian distribution of text types). One solution to this is to use stratified sampling[150], allowing for control over the sample proportions without entirely sacrificing representativeness. Biber's presentation of this idea has seemingly led many to reject the common wisdom of sociological sampling, leading to corpus building to be seen as a fundamentally new activity, and consequently as something of a black art.

A further aspect muddying the waters of quantitative representativeness assessment is disagreement over how to parsimoniously stratify language. This is in part due to the difficulty in defining a taxonomy for genre[110], which is often the most controlled variable within a general purpose corpus' sample design. Ultimately, I consider that the answer to this is, as mentioned in the LOB manual, down to the individual research question: a corpus sufficient to represent one feature may easily be massively biased for those with greater variation in the population.

In part for the reason that Biber's work was taken early on as a validator of current practices in corpus building, the notion of representativeness has remained an almost entirely philosophical concept within corpus linguistics.

The BNC, though often relied upon as a reference corpus, makes a number of bold assumptions regarding representativeness—to the point of explicitly stating its deviation from being a solid representation of its population [31, p.6]:

> *There is a broad consensus among the participants in the project and among corpus linguists that a general-purpose corpus of the English language would ideally contain a high proportion of spoken language in relation to written texts. However, it is significantly more expensive to record and transcribe natural speech than to acquire written text in computer-readable form. Consequently the spoken component of the BNC constitutes approximately 10 per cent (10 million words) of the total and the written component 90 per cent (90 million words). These were agreed to be realistic targets, given the constraints of time and budget, yet large enough to yield valuable empirical statistical data about spoken English.*

This implies that the corpus should not be used as a single sample at all, or that large adjustments should be made during analysis to avoid generalisations on purely quantitative bounds.

A similarly pragmatic approach is taken to the sampling of written material according to production and reception statistics, with a balance being brought between the two based on publication, library lending, and magazine circulation statistics. Further, some samples were taken purposively [31, p.10]:

> *Half of the books in the 'Books and Periodicals' class were selected at random from Whitaker's Books in Print 1992. This was to provide a control group to validate the categories used in the other method of selection: the random selection disregarded Domain and Time, but texts selected by this method were classified according to these other features after selection.*

The spoken portion contains a 'demographically sampled' portion (comprising 50%), so called because it makes an attempt to represent speakers according to their sex, age, and social class. Individuals were randomly distributed, and each provided recordings of their conversations over a two to seven day period. The major limitation noted here is the short period of time, and thus the low probability of capturing certain, rare, interactions on tape. In order to remedy this, half of the spoken component was devoted to genre-based ('context-governed') stratification, with the intent being to capture a greater breadth of data.

By contrast to corpus linguistics, NLP has written little on representativeness, focusing more on the parsimony and utility of models than their external validity. A notable exception is the test by Dridharan and Murphy[169], who attempt to assess the impact of corpus quality on distributional semantic models. An illustration of the conceptual split here may be seen in their use of the term 'quality' to refer primarily to the form of the data, and the degree to which it is curated manually. This is of particular relevance as their corpora are sourced online, something discussed in Section 2.3.

Size

The size of any sample is a crucial and often hard-to-determine property, and corpora do not differ in this respect. In many ways the question of corpus size is a primary constituent of the representativeness property above, and though it has long been recognised as such there remains little consensus on just how large is large enough.

This disagreement is in part because language exhibits a number of properties that make it hard for us to gauge variability in the population, meaning that most sample size estimation methods (which rely on random sampling) are poorly suited.

The first of these is the inability to accurately measure the complexity of language, which, as used and applied by humans, has unknown degrees of freedom. This effect applies itself at many levels:

- Lexicon—Vocabulary, even when restricted to a given demographic, time period or person, is difficult to define with any certainty. Many people are capable of recognising entirely novel words, simply by virtue of their context or morphology (and the inverse, e.g. *Jabberwocky*).

- Syntax—The meaning of words and phrases is heavily context based, but the context affecting each aspect of language is variable and, in some cases, wide-reaching. This makes it hard to determine if we should be sampling single sentences, 2,000-word texts, or the whole thing[75]. This unknown sampling unit drives one of the key trade-offs in corpus design, as a 1-million word corpus comprised of single sentences will be capable of embodying more variation than will one made of two 500,000-word samples.

- Semantics—Variation in our understanding of language in context is poorly understood. This is the driver behind any models of the above, but also affects how we should sample external variables such as socioeconomic factors and even the direct circumstances in which a text is used[165, 171].

This ambiguity produces a tradeoff that has been identified by relatively few in the corpus-building community[46, 94, 67, 8]—that corpora of equivalent size may yield significantly different inferential power due to their differing number of data points. This may be seen as an issue of depth vs. coverage: corpora including large snippets of text are capable of supporting more complex models and deeper inference, at a cost to the generality of their results.

There remains disagreement upon the extent to which these two aspects should be traded off, though it is notable that the NLP and linguistics fields vary greatly in their treatment of the data, with linguistics typically focusing on frequency and immediate collocation, and NLP being skewed towards more complex, instrumentalist, models. I would suggest that, realistically speaking, researchers should be examining their experimental design with respect to the size and/or complexity of the features they are working with in order to select a corpus that matches most closely. It is also quite clear that this does not happen: the BNC contains a small number of large samples, yet is often used to analyse small-scale linguistic features.

Secondly, establishing the boundaries of a given language is difficult. Users of a general-purpose corpus wish for two conflicting properties to be satisfied—the population must be wide enough to provide a useful set of persons about which to infer (and, more loosely, this should align with those persons we can say informally, for example, 'speak English'), yet the corpus must provide sufficient coverage to represent that population in the first place.

Generally it seems that this problem, though having been recognised, has received too little effort for practical reasons. Issues of balancing demographics in corpus design have typically taken a curiously detached form, that of selecting a sample of language as a proxy for demographics (e.g. selecting the bestsellers over unpopular books or sampling more popular newspapers).

This approach has led to a situation where each and every general-purpose corpus carries with it the expert judgement of linguists not only in selecting a wide variety of texts from within the population, but also their socio-linguistic opinions. Sinclair makes this explicit in 'Developing Linguistic Corpora', where representativeness is described in entirely subjective terms [185, p.5]:

> *A corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it, so users must be able to refer to its make-up in order to interpret results accurately. In everything to do with criteria, this point about documentation is crucial. So many of our decisions are subjective that it is essential that a user can inspect not only the contents of a corpus but the reasons that the contents are as they are.*

This reliance on expert opinion to overcome practical challenges associated with text retrieval has lead to the sampling policy being somewhat opaque to end users. Those using a given corpus are not necessarily able to rigorously define about whom they infer a given result. In practice, this manifests as a need to qualify results by reasoning about the likely impact of any ambiguity.

Finally, disagreement on how we should extract features from language samples (i.e. which dimensions of variability are interesting) means that, aside from the immediate and obvious properties such as genre (about which there is arguably less agreement[110, 6, 161]), any efforts to stratify language are met with suspicion. This may lead to researchers performing corpus analysis using informally-subsampled general-purpose corpora, with questionable correspondence between the categories used to select texts and their research goals. It is the author's opinion that this is unanswerable except for individual studies: in order to know variation in features affecting one's use of a corpus, it is necessary to define the covariates and evaluation strategy.

Further to these problems of defining the nature of the sample, there is the resultant problem of determining a sample size even where these are known.

Taking the first of these issues in the extreme, it may be said that the only 'fully representative' corpus must contain all context for each text (something that would include at best an abridged history of the world) in order to satisfy inquiry from many different perspectives. Given the limitations in analysing properties of language beyond a given scale, and the clear impracticality of extending that scale's upper bound, it seems reasonable to conclude that current corpus

efforts are sized so as to be useful for small-scale features, and that our inspection of language is currently somewhat shallow.

Taking into consideration problems of proportional, stratified sampling, it seems possible to establish a corpus size and composition that is widely agreed upon. Nonetheless, issues of selecting a sampling unit mean that the resultant corpus may either be a refinement of current efforts (not necessarily a bad thing) or utterly colossal and beyond the capacity of even modern corpus processing systems.

Since sample size and composition are intimately related both to one another and to the concepts of representativeness and transferability, this issue will be one of primary importance to the rest of the thesis.

Purpose

The reason for sampling a given population is a crucial feature of corpora. Not only does it define the level and type of metadata available (and the arbitrary definitions used therein), the expert selection of sample designs means it has a large influence on the sampling frame used, and thus the validity of any generalisations made using said data.

The condition given by McEnery & Wilson[127] here is that a corpus must have been built with some degree of linguistic inquiry in mind, for example, the collected works of Shakespeare would be counted, but not one's personal book collection (even if it happens to include the complete works of Shakespeare and nothing else). This distinction seems to be little more than a way of stating that one's ability to infer things from a corpus is relative to its external properties, which is true of any sample.

This requirement calls into question two things: the generality of a corpus, and the extent to which it is documented.

General-purpose corpora sample a large population, of interest to many users. This requires that they remain fairly unbiased and cover a large amount of variation (which in turn necessitates very large sample sizes). Since they are re-used many times, the quality of their documentation is key to their gross value—each user will require particular variables in order to generalise from the text.

Special purpose corpora avoid this challenge by being re-used for less disparate aims (and, generally less frequently). Because of the relative focus, their documentation is often able to be significantly more detailed, allowing for deeper insight. This approach, however, is not transferable to the sample sizes used in general purpose corpora.

To contextualise this, both examples above merely constitute special-purpose corpora, that offer answers to different research questions. For one, we may answer those about the nature of Shakespeare's language use, whereas the other allows us to find knowledge about a given person's literary preferences. It is of no direct consequence that more people are likely to care about the former.

If we extend the example to include others' book collections, the ability to generalise changes: we know significantly more about Shakespeare than many other authors, and it is thus possible

to annotate the Shakespeare corpus with details of his life, times when works were written etc. The same could not be said of a corpus where our aim is to investigate reading habits (even if some people exclusively read Shakespeare).

This 'purpose' requirement can be phrased entirely in terms of external variables. A group of texts about which nobody knows anything do not offer any opportunity for inquiry (except about themselves), and so it is reasonable to require documentation of the context in which those texts occur (or any external property that demarcates a homogeneous group).

By stating that a corpus must have a defined purpose, we include a set of reasonable assumptions about the corpus and its external properties: a corpus built for study of Italian newspaper text is unlikely to heavily sample The Guardian, for example. In many ways, statement of purpose offers a shorthand for many decisions and assumptions inherent in construction of a large corpus, and a simple way to assess how well matched a corpus may be for another—related—purpose.

It is noteworthy, however, that selecting a corpus by its original purpose greatly complicates any subsampling that is possible—one must be able to subsample texts not only according to external data of interest, but also taking into account the original interactions and assumptions made by the constructors. As mentioned above, it is unlikely that any corpus is capable of being documented 'fully' enough to avoid these issues, however, this is a compelling argument for focus on detailed metadata being provided (rather than detailed rationales for sampling), especially in the case of general-purpose corpora—say 'what', rather than 'why'.

Data Format

Some authors stipulate that a modern corpus should be electronic. To generalise this position, they require that it is in some way processable by machines, i.e. that it must be in some way regular. This defines the format of not only the basic textual content, but also the availability of metadata at all levels (category, document, word).

Both of these have been addressed to some extent, especially the problem of representing the text itself, which has largely been solved by UTF-8 at the character level, and XML/SGML at the markup level. Standards derived from efforts such as the TEI[80] have some penetration, though there is often still the need to include proprietary extensions if other concerns on this list are to be maintained.

Though earlier work on corpus construction focused on data formats[7, 166], with a view to sharing corpora for local analysis, modern approaches tend towards corpora 'as a service'[71, 49]—providing a front-end to query and analyse text directly whilst still hosting it at the original institution. This approach is taken to mitigate licensing issues, and to work around the high level of technical skill necessary to work with large-scale data.

This thesis takes the view that representation is largely a solved problem: UTF-8 and XML are both capable of storing international characters and complex annotations in a way that is easily mined for many uses, and advances in database technology and distributed processing offer many ways to process and retrieve structured data.

Classification

Efforts have been made to address the ambiguity of some often used strata such as text-type and genre. One of the more notable was EAGLES[166], which produced a number of recommendations designed to be applied to new corpora whilst retaining general compatibility with existing designs.

To some extent, the form a corpus takes is defined by its content—any meaningful taxonomy is arbitrary and theory-laden. For this reason, recommendations made by EAGLES were driven by a review of the theoretical basis for existing taxonomies and work at the time[166]:

> *Any recommendation for harmonisation must take into account the needs and nature of the different major contemporary theories.*

The need to provide useful metadata is particularly challenging for general-purpose corpora, which have very loosely-defined aims and must maintain a high level of neutrality.

Internal text distinctions may be drawn using 'bottom-up' (often termed 'corpus-driven') methods, however, these are frequently difficult to operationalise. A prime example of this is Biber's multidimensional analyses of corpora[21], which focus on feature extraction and principal component analysis in order to identify primary dimensions of variation within corpora.

Biber's approach has been held aloft by many as one of the only 'unbiased' analyses of variation around, however, this is not the case. Without an authoritative underlying theory of language use by humans, any features used to construct the model will, however neutrally treated thereafter, exert pressure on the results. Given the subtlety with which such analyses extract information (and the difficulty in interpreting resultant factors), this is likely to lead to favouring one set of conclusions over another.

This is also true of higher-level techniques such as latent dirichlet allocation (LDA), which is usually trained using token frequencies[25]. These straight frequencies are still entirely bereft of context, and tokenised using procedures that embody particular theoretical decisions such as the importance of punctuation[138] (especially true for languages lacking word delimiters, such as Chinese[184]).

Sharoff[160] presents a more transparent clustering methodology that relies on keyword metrics to identify salient features. This provides some connection to other keyword literature, along with a mechanism for inspecting the resultant categories.

Arguably, further problems with the approach of finding natural strata of variation lie with sampling and linguistic problems: many of the practical issues surrounding corpus building involve enumeration of easily identifiable groups of texts, something that would be hard to compromise or 'trade off' if working from factor loadings. Further, many analyses will find subsampling data difficult when defined in terms of many external variables[6], though this is more of a challenge for the linguistic community (and providers of its tools).

Through reasoned examination of existing efforts, Lee[111] was able to re-form the BNC's classification by adding external data and classifying documents manually. This is arguably the most prominent effort to apply the multi-phase 'examination and re-appropriation' method that

Biber and EAGLES recommend.

His methodology, though labour-intensive, offers a defensible way to trade off the various interests of users. One key aspect to this is the fact that it was built after the corpus, and thus may take into account common usage when making distinctions between texts—something that Lee relies on in an effort to define categories using 'external' definitions (as opposed to Biber's internal variance measures). This may prove more easy to operationalise in some circumstances, but still involves the subjective expert opinion of someone who may or may not agree with the user's perspective[6].

In summary, the problem of producing a meaningful and operationalisable taxonomy for corpus organisation is actually one of community agreement: for any given user of a corpus, the task is simpler. This indicates that a transparent, well-documented approach should be taken in order to allow end users to decide upon their level of agreement with the pre-applied categories, or that differing levels of confidence should be applied to aid subsampling.

Dissemination and Collaboration

The ability for multiple researchers to access a corpus is one of the main benefits of corpus methods—corpus-based studies may be replicated and compared with absolute certainty of the empirical aspect of the research.

The process of building a large, multi-purpose sample for use by a whole field of research mirrors that used in other fields, many of which have similar problems gathering data.

Many examples of this approach exist, such as the British Household Panel Survey and British Crime Survey[173, 76], both of which demand[ed] significant investment over a long period in order to overcome the practical issues of large-scale demographic sampling. Nonetheless, the quality this yields has lead to their widespread use, for example, both are used extensively by governments to assess social policy impact.

Although corpora are generally less sensitive on ethical grounds, many legal challenges remain to distributing and using such a large quantity of data. This has historically greatly limited both the source and form of data gathered for corpora, and was one of the reasons behind sampling 2000-word samples (rather than whole texts) in the Brown corpus.

Navigating copyright law remains one of the primary tasks of a corpus building effort, though the increasing dependence on digital sources has dulled this somewhat, since they are typically less controlled when being published. Nonetheless, corpora often come bundled with restrictive licensing, something that limits the ability of the wider community to participate in their use.

Many countries' fair use exemptions apply to research, though this may apply only to copies taken for private study, as laid out in the UK's Copyright, Designs and Patents Act 1988 . The concept of 'Fair Dealing' covers many uses of extracts in research, and a recent exemption was added that additionally allows the use of[178]:

> *. . . automated analytical techniques to analyse text and data for patterns, trends and other useful information.*

The heavy re-use of corpora, then, exerts both positive and negative forces upon the scientific community. On one hand, it is necessary to share resources in order to lower costs, increase awareness of rare data, and pool efforts to create better samples. On the other, the ubiquity of a corpus may damage its scientific value, and starve the community of more up-to-date or relevant resources. In the worst case, widespread use of a single corpus by the community may lead to partial circularity of hypothesis derivation and testing.

In an ideal world, it would be possible to replicate studies with one's own samples. As I shall cover throughout this thesis, increasing digitisation of documents (and use of the web) offers a way to make this scenario possible, at least for some corpora and purposes.

Sample Type

Corpora span many types of sample, from simple cross-sectional ones ('synchronic' corpora: Brown, BNC) to those that aim to report language use through time ('diachronic' corpora: ICE, Longman/Lancaster) and a hybrid of the two, which is designed to follow language use and update on-the-fly ('monitor' corpora: BoE, COCA)[52, 30, 62, 172, 85, 37]. Further to this, there are parallel corpora, intended to match texts across some variable such as language[102, 125], and many other designs that combine properties of these to satisfy various sample designs.

These approaches represent various use-cases, and are often significantly less 'general-purpose' than large synchronic reference corpora such as the BNC. This thesis treats the selection of a sample design as problem-dependent.

A Working Definition

The above discussion illustrates the wide variety of samples that may be called general-purpose corpora, and some of the issues that affect their utility. The focus of this thesis is on mechanisms for making, operationalising and documenting the above choices, and as such the working definition here opts to not apply any clear restrictions.

One obvious requirement, however, is that a corpus must be a suitable sample *for its intended purpose*. This means that the sample design (external variables) and the document classification (internal variables) must be clearly defined in terms of the research questions given.

For the purposes of this thesis, a corpus constitutes a body of text that must be:

- Representative of some stated population;
- Sufficiently large to satisfy that representation (for a given set of purposes);
- Explicit in its coverage of external variables;
- Of a regular, machine-readable format.

2.2 Sampling Principles

Corpus building methods are largely based on sampling methods from the social sciences. These methods are well developed, and their formal frameworks specify a number of design choices

that must be made whilst designing a sample. These decisions largely affect the suitability of a corpus for different forms of analysis, and the frameworks they are based on may be used to motivate design of the sampling process itself. Re-examining the original principles of sample selection allows us to use some of the formalisms developed elsewhere to inform judgements on the properties of linguistic samples.

The goal of any sample is to present a scaled-down set, containing individuals that represent all variation within the population. Before discussing the implication of various approaches, it's therefore important to draw a distinction between variables which are controlled by an experimenter (independent) and those that remain free to vary (dependent).

This distinction has a large impact on sample design, as it is impossible to draw conclusions about the population by inspecting independent variables. Further, the selection of documents according to these controls often leads to systematic bias in dependent variables. In order to come to conclusions about representativeness and sample quality, it is thus necessary to identify these variables ahead of time.

In the case of sampling documents, I will be presuming that users of corpora are primarily interested in inspecting 'internal' text features, which are described in terms of 'external' document metadata. This guiding principle mirrors the metadata/data dichotomy seen in corpus tools, and should be uncontroversial[2]. In this thesis, I will often label variables as internal or external based on these criteria, with the implication that external variables are independent, and vice-versa.

The ideal sampling scheme for a given population and selection of variables, then, contains variables as controlled specifically for the research question about which one wishes to infer, and maximises coverage of internal document content (and uncontrolled-for metadata values).

Taking this principle to extremes yields the maximally-representative census sample: 100% of the population of interest. At this point, any inference is mere observation, and the only potential pitfalls are ones related to whether or not the question itself is worth asking, rather than the validity of its answer. A census still contains theory-laden assertions, however, in the form of its population definition[7].

At the high level, samples may be classified into two main groups: *nonprobability*, and *probability* samples[135]. The former of these is primarily guided by a systematic or subjective choice, and the latter has a sampling frame defined by random selection.

2.2.1 Nonprobability Sampling

The selection of data points in a nonprobability sample is performed either by an objective system, subjective reasoning, or some combination of the two. Factors influencing selection are often situational or theoretical, meaning that samples require a greater understanding of the subject to avoid accidental biases.

Three main forms of nonprobability sampling are identified by Barnett[12] and Teddlie[174]: *availability sampling*, *purposive sampling*, and *quota sampling*.

[2]Note that many definitions of genre, text type, etc. make this circular, as they are defined by document content.

Availability Sampling

Also known as convenience sampling, this approach simply takes the most readily available data points.

This method has been widely used in the social sciences, with experimenters often using students from their affiliated institutions (an approach used for some very famous studies).

Convenience sampling is rightly regarded as extremely unrepresentative of wider populations, particularly in fields (such as psychology) where there are a great number of covariates. If corpus linguists were to work with data sampled as conveniently as those in the Stanford prison experiment[3][70], they would merely be reading books from their own bookshelves.

There are a number of methods that are designed to adjust for these biases. Birnbaum & Munroe[24] present a model of the bias resulting from unavailable population members after random selection. This approach is difficult to operationalise in a linguistic context as they require enumeration of data points prior to determining whether or not they are available, rather than the more haphazard approach of true convenience sampling.

Farrokhi[48] approaches the problem of constructing two groups: a control and treatment group, using a series of predetermined criteria to assign membership. This approach mirrors the comparative nature of many corpus linguistic methods, but does not directly offer a way to produce a more representative corpus for 'general use' aside from the principles of using heuristics to guide design.

Despite the drawbacks of convenience sampling, such an approach may be appropriate for populations where it is agreed that there is little variation between individuals for the variables of interest: for example, a study which seeks to establish the modal number of eyes humans have might be well served with a very small sample. Smaller linguistic features such as unigram frequencies are likely to be similarly easy to represent.

Purposive Sampling

Purposive sampling describes the case where those constructing the sample make a deliberate and systematic choice of inclusion, based on expert opinion.

This approach is primarily effective against well-known sources of bias, and the validity of any sample built using it is entirely dependent on identification of these. Where the expert design encompasses all variables of interest to a study, and where selection has been performed in such a way as to encompass individual data points from across the population, such an approach may result in a high-quality and defensible data set.

The main difficulty is in avoiding previously-unknown correlations between the theoretical selection criteria and the study variables. As selection criteria are based on domain knowledge, and heavily theory-laden, it is often difficult to anticipate their interactions with other variables of interest. As theories improve over time, this may also to lead to the effect of samples being considered more biased as knowledge of an area improves, gradually reducing the perceived

[3] "The participants were respondents to a newspaper advertisement, which asked for male volunteers to participate in a psychological study of 'prison life' in return for payment of $15 per day."

validity of any results.

Criteria for selection generally accomplish differing goals, and will suit varying study aims[38] (for example, some of these may be very useful for exploring rare features):

- **Heterogeneous**
  An attempt to cover as much of the population as possible, by selecting data points that are unlike ones already in the sample.

- **Homogenous**
  Data points are selected to be as similar as possible according to given variables. This is suited to in-depth qualitative analysis, being somewhat analogous to merely controlling for more variables (i.e. reducing the population).

- **Typical Case**
  Data points are selected according to a theoretical/rational idea of typicality.

- **Extreme Case**
  Only those data points regarded as atypical are sampled. May be used to explore reasons for atypicality.

- **Critical Case**
  Data points are selected based on previous theories, such that they explain certain hypotheses.

- **Total Population**
  An entire sub-population is sampled, due to its ease of access (for example, all members of an organisation, or all publications by a certain author). If no other items are sampled, this merely becomes a census with a very tight population.

- **Expert**
  Expert opinion is used to determine inclusion in the sample.

From a corpus construction perspective, where the sampling party is often distinct from the analyst, the complexity of sample inclusion criteria brings with it the need for extensive documentation: without awareness of the expert's choices, it becomes very difficult to defend any use of the sample. Due to the nuanced nature of their validity, purposive samples are valid only for qualitative analyses, where interactions with the study design are able to be rationalised and explained in context. This is especially the case where a study's purposive design is borrowed for use in other contexts: for example, the Brown corpus explicitly states its aim to represent 'standard' English, yet its sample design has been widely copied[78, 162, 128].

In fields which lack a cohesive, quantitative model of their population, it is often necessary to start from an expert-defined base. Ideally, information from this sample may then yield methodological improvements, and, eventually, an unbiased mechanism for retrieving a statistically representative sample.

Quota Sampling

This is a two-stage design, with a number of sub-populations being identified and then selected on a purposive/availability basis. It is essentially a nonprobabilistic form of stratified sampling: the population is split into mutually exclusive groups, into which data points are placed until each group is 'full'.

This approach is often used to control convenience sampling variation for some important variables, for example, controlling for sex and age whilst performing market research. The BNC's spoken portion was 'balanced' using this method, with a number of bins being allocated by context and speaker information. The COCA corpus was also constructed using this method using data from the web [36, p. 163]:

> *Using VB.NET (a programming interface and language), we created a script that would check our database to see what sources to query (a particular magazine, academic journal, newspaper, TV transcript, etc.) and how many words we needed from that source for a given year.*

Summary of Nonprobability Methods

Corpus sampling methods described above already closely resemble quota sampling, using a lot of expert opinion to define the quotas. Nonprobability sampling, however, is particularly ill suited to statistical analysis and inference, which relies on random selection over uncontrolled variables. Simply, nonprobability sampling techniques are very easy to bias [12, p.19]:

> *. . . there is no yardstick against which to measure 'representativeness' or to assess the propriety or accuracy of estimators based on such a sampling principle.*

This opacity leads to limitations in corpus analysis, where the goal is to use large volumes of data as objectively as possible. For any quantitative analysis to be scientifically defensible on empirical (rather than rational) grounds, probability sampling is *required*.

2.2.2 Probability Sampling

There are many probability-based designs available, and the choice of them is largely dependent on the methods of inquiry that are to be applied to the resulting data. In all cases, the goal is to allow the dependent variables to vary randomly, ultimately allowing for statistical inference if sample sizes are sufficient.

Random selection of data points provides both a mechanism for unbiased estimation of parameters such as means, as well as knowledge of variation. The latter of these is the key difference between a probability-based sample and a well-chosen nonprobability one, and is vital to basic hypothesis testing procedures. Further, notions of sample size are entirely based on these measures: power analysis demands some knowledge of variance and effect size, both of which are only meaningful under conditions of random selection.

Random sampling methods identified by Lohr, Barnett, and Teddlie[118, 12, 174] include: *simple random sampling*, *stratified sampling*, and *multi-stage sampling*.

Simple Random Sampling

This approach selects members of a population entirely at random. Each member of the population has a probability of selection that is simply $\frac{sample\ size}{population\ size}$.

SRS is free of any errors that stem from classification, reducing the importance of potentially circular genre definitions and taxonomies. The main disadvantage is that it requires a complete sampling frame: that is, all individuals in the population must be known in order to be randomly selected from.

Though corpora often contain randomly sampled portions, for example, where data is enumerated by a publisher's list or online directory, the lack of a central authoritative index that covers the whole population is usually an impediment to retrieval of a truly representative sample. Essentially, simple random sampling is unsuitable for larger samples in linguistics due to this limitation.

Once a corpus has been constructed, it is often possible to use SRS in subsampling approaches. Its unbiased properties also make it useful in bootstrapping, allowing methods to use the full distributional information contained within the corpus[64]. These techniques serve to avoid introduction of further bias, but don't sidestep any representativeness issues with the original sample.

Stratified Sampling

Stratification is the process of breaking a random sample into a set of bins, with sizes weighted according to some policy. In the case that the strata are selected in an unbiased manner, this should yield the same sample as SRS.

This method can be used to ensure that subpopulations are sampled comprehensively by overestimating their proportions, or by using a different sampling method for that stratum. (for example, selecting more people from areas with low population density). Such practices damage representativeness but may be useful for qualitative analysis[175].

Stratum selection should be along real-world subpopulation boundaries, and is usually selected in order to maximise either representativeness (by ensuring that strata are sized according to the population) or statistical power (by ensuring that strata are sized according to the amount of variance in the strata). Stratification is most efficient when strata are homogeneous and well separated.

Stratification requires the ability to exhaustively separate the population into mutually exclusive categories, and sample from these categories in a random manner. Both of these are a challenge in corpus linguistics, as removing the first level of 'no indexing' often leads to another. Nonetheless, where information silos exist (such as in academic publishing) then this approach is particularly suited[150].

Where auxiliary data exists, statistical benchmarking may be performed by adjusting sample weights according to the proportions seen in this data. This may be applied after sampling itself (otherwise said auxiliary data is simply used to determine the initial strata sizes), and so is often referred to as 'post-stratification'. This approach is applicable where ordinary stratification is impossible, for example, where the variables on which to stratify are unknowable at the time of sampling.

Multi-stage Sampling

Multi-stage sampling involves multiple rounds of random sampling with progressively diminishing sampling units.

Multi-stage sampling is particularly applicable in cases where *all* of the data points in the population are accessible through a hierarchical structure. This is clearly the case for smaller linguistic features, but much less so for whole documents (or large extracts).

This approach often significantly speeds up the process of enumerating and selecting from a population, and reduces the need to classify things compared to stratified approaches. The increased structure may also be of use to some models, allowing for multi-level modelling approaches during analysis. Because the method relies on excluding large areas at once, it is less representative than SRS for the same sample size, and this may complicate analysis.

Cluster sampling, a form of multi-stage sampling that relies on existing organisation of data, has particular applicability to web sampling, where documents are often stored in academic repositories or simply under the same domain, or to sampling from different publishing houses.

Evert provides a library metaphor for sampling in which he explains random selection of individual words using a multi-stage approach[46], selecting progressively smaller units at random before returning a single word and repeating the process. Though he focuses on randomness, it would be possible to use this approach with stratification at each level—this mechanism is well suited to situations where auxiliary data may be available, but not at the level of the desired sampling unit.

Probability Sampling Summary

Simple random sampling is often seen as the ideal probability sampling design, in that it yields high quality results using statistical methods, without the need for weighting and adjustment of results. It is also by far the hardest method to apply to text due to its incompatibility with the process of accessing the data.

Stratified sampling is arguably simpler, in that it allows for testing and definition of strata prior to retrieval of texts, and stratum sizes may be computed from existing corpora in a multi-stage design. Another benefit of stratified approaches is the ability to examine and adjust the distribution of strata according to expert opinion, offering a hybrid design that is able to compensate for known practical issues. This is often used to artificially boost the stratum sizes of minor groups within the population in order to ensure that they are over-represented in the

final sample — something that may be desirable if their influence is particularly important to a research question.

2.2.3 Sample Size Estimation

Quantitative sample size estimation is largely based on ensuring that a given hypothesis test can detect an effect of interest. This varies according to a number of parameters, including the type of test used. Any hypothesis test is defined by four parameters[44]:

- **Probability of Type I Error (α)**
  The probability of rejecting the null hypothesis when no effect exists in the population. Forms the threshold for the oft-quoted 'p-value'.

- **Power ($1 - \beta$)**
  The probability of rejecting the null hypothesis when an effect exists in the population.

- **Effect Size (d)**
  The observed change in a parameter necessary to identify an effect with probability $1 - \beta$.

- **Sample Size (n)**
  The number of individual units in the sample that are free to vary.

Sample size must be estimated with knowledge of the power of the test to be done, and of the population distribution. A simple binomial test may have its sample size estimated thus:

$$n = \frac{Z^2 p(1-p)}{d^2}$$

Where Z is selected according to the area under the population distribution according to α, and p is the parameter value within the population (often estimated as $\hat{p}$). This implies that a reduction in any effect size we wish to detect increases the required sample size (and vice versa), as do increases in the confidence at which we seek to reject H_0 at. This general trend applies to all sample types above.

In addition to sample size, the power of a test is also affected by the method used to perform the hypothesis test. Tests which are excessively conservative further increase the probability of missing an effect, lowering power, as does violation of assumptions upon which parameter estimates are based.

Current methods for computing sample size are reliant on assumptions of independence, often phrased such that members of the sample must be 'independent and identically distributed' (IID). This assumption is violated massively by conventional corpus construction: corpora are sampled at an extract or document level, yet often analysed at a word or phrase level.

In the absence of IID samples, sample size calculations are reduced to what Cohen [34, p. 145] termed 'non-rational bases' such as past experience or rules-of-thumb. The process of building a corpus that is IID for a given research question is so theory-laden as to be inapplicable to those building general-purpose corpora using the current model (where many users share a large corpus for diverse tasks).

The reliance of corpus linguistics (and to some degree NLP) on qualitative methods and rational definitions of parameters such as population and sampling frame make application of existing power analysis to corpus construction challenging.

An alternative method, suited to qualitative and mixed-methods approaches (and thus less reliant on the randomness assumptions violated by document-level sampling) is presented by Glaser[58], who suggests that those performing experiments should be guided by 'saturation': the point at which no (or negligible) new information is presented by adding new data points to the sample. This mirrors the methods of Good-Turing frequency estimation[61], as used by Biber in his 1993 assessment of representativeness[22].

Saturation-based measures offer a less-formal mechanism for assessing the adequacy of sample sizes, however, they still require definition of a study design prior to assessment of sample size sufficiency. Such methods are also particularly lacking when sampling is nonrandom, as they rely on the discovery rate of new information being unbiased—if retrieval of documents is systematic the rate of new information will plateau, leading to premature conclusions of coverage. Again, without the ability to enumerate the population, it is difficult to insure against such bias.

2.2.4 Sample Design

The design of a sample can be broken into a number of stages, each of which informs the next. As we have seen, many of these require theoretical justification or analysis based on the final study design. This forms the framework used elsewhere in the thesis as an ideal case.

Research Question Definition

The definition of a research question is an important first stage to selecting a sample. This is in order to define any theoretical assumptions and the analysis design, both of which have significant impact on the size and form of any sample.

Sample Unit Specification

The unit of analysis should be in agreement with the aims of the research question above, and ideally should match the sampling unit. Where disagreement occurs, this is likely to abrogate the quality of any analysis based on IID assumptions, including power analysis for sample size estimation below.

Variable Selection

Dependent and independent variables should be specified unambiguously, and any relationships between the two should be examined from past experience and existing literature, in order to reduce unanticipated correlations causing bias. Where the research question and intended analysis are already rigorously designed, this task should be fairly transparent to any circularity, however, this becomes more challenging as qualitative elements are included.

Population Definition

The bounds of the sampling frame must be defined in terms of the external variables to be sampled, such that any retrieval efforts can unambiguously use these. A population definition is, in effect, definition of an independent variable which is controlled prior to sampling, and as such its relationship to dependent variables and to the original research question is similar to that of any other variable.

Sample Policy

The manner of sample must be decided based on practical issues, and on the various properties specified above.

Sample Size

Depending on the design and analysis method proposed, this may take the form of a quantitative power analysis, or a plan to implement qualitative controls during construction of the sample. Some a-priori objective criteria should be defined, particularly when a sample is based on qualitative assessment, in order to avoid data dredging (or 'dredging' of variables correlated with those to be studied).

The existence of previous studies, literature, and datasets may guide all of the above stages, and the progression from expert-opinion-based sampling to a more quantitative understanding of population variance is made possible by iterating the above: something that happens naturally as a field progresses.

It is notable that traditional general-purpose corpus construction efforts are largely unable to specify research questions ahead of time as they are necessarily retrieving data prior to the study even being conceived. It is also the case that most studies include a significant qualitative component, in interpreting the output of quantitative models or in the form of direct reading[146]. It seems likely that such properties pose the biggest challenge to the classic question "how large should a corpus be?", which requires more rigorous specification of these.

2.2.5 Sources of Error in Corpus Construction

The validity of a sample is based not only on its representativeness for a given question, but also how well that question may be related to reality, and what limitations are imposed by the process of retrieving data. Corpus builders are not blind to these challenges—indeed, most literature on corpus construction devotes a large portion of its content to practical issues[185, 7, 166].

The requirements discussed above are satisfiable in a number of ways, each of which will exclude and promote certain uses of the result. Generally, threats to the validity of inquiry based upon these samples may be broken down into three areas:

- **External Validity**
  How relevant the sample is to the population about which we wish to infer something.

These are likely to limit the generality of a study, or lead to under/overestimation of its effects.

- **Internal Validity**
  How much a study can rely on document annotations and data in order to draw conclusions. Issues here are likely to cause false results.

- **Practical Issues**
  Limitations on the mechanics of sampling. These issues may cause either or both of the above.

This section identifies potential issues with corpus sampling methods. The majority of these are practical issues for which fixes must be designed carefully and on a case-by-case basis.

Distributional Issues

This category largely describes the manner in which certain important covariates are selected for sampling, or sampled. Language contains an unknown (but undoubtedly very large) number of possible dimensions to study and compare, and selection of features to describe variation across texts is far from simple. Such covariates may be listed as external metadata (author age, genre) or for linguistic features (prepositions, wh-relative passives).

Atkins et al.[7] provide a series of recommendations for covariate selection, which has been used to guide this discussion.

The population for whom a corpus describes language use is relatively difficult to describe, as it is largely a self-referential problem. Seemingly, one reason why corpora are not demographically representative is because their selection processes rely on lists that are compiled using data for which it is difficult to determine comparable bounds. The BNC's policy for selecting written materials, for example, is segmented into the following sources:

- Books, selected from bestsellers, literary prices, library loans, additional texts;
- Periodicals and magazines (including newspapers);
- Other media (essays, speeches, letters, internal memoranda).

Of the lists used in the books category, they specify the following criteria [30, p. 9]:

> *Each text randomly chosen was accepted only if it fulfilled certain criteria: it had to be published by a British publisher, contain sufficient pages of text to make its incorporation worthwhile, consist mainly of written text, fall within the designated time limits, and cost less than a set price.*

It is often difficult, using other sources of data, to establish the details of an author's nationality, the age of a text, or the suitability of the time limits[40]. These issues are typically better addressed by more specialist corpora, which are subject to easier-to-determine bounds such as social role, context or text type[104, 142, 107].

Ambiguity in population definition has a number of direct influences on other issues of corpus validity. Selection of stratum sizes, for example, must be based on the relative proportions of language used by the given population. If a population is defined using purely internal (linguistic) properties, this becomes a reflexive and circular task—we end up selecting proportions of language to match proportions seen in language. Use of auxiliary data to augment sampling policies (such as social demographics taken from other, large-scale surveys) must also be matched to the population in question.

Where selection of genres is defined by linguistic content, it is necessary to ensure that the frequencies used do not have systematic correlations with features being studied. This is impossible to automate, and must be assumed based on experience and theoretical reasoning.

Speech corpora are often specified in a significantly less ambiguous manner (though not always with more proportional sampling). This seems to be due to the direct nature of speech sampling, which involves the person actually performing an utterance (rather than the language use itself being a persistent and concrete artefact).

This difference is identified by Leech[115], and will be inspected in greater detail below, as it is a potential major source of disagreement between construction and use.

One source of ambiguity when defining a population is inherent 'measurement error' in the classification scheme used to delineate the texts. Efforts such as EAGLES[166] identified this issue as a significant one, and much effort has since been spent on minimising the problem of classifying texts. Nonetheless, the difficulty in identifying a widely-agreed-upon classification scheme means that this remains an important design decision.

This manifests in two ways. The former being the problem of selecting texts according to external variables about which we may wish to infer findings, such as level of education of the author, nationality, target audience of text, etc. Many of these factors are imposed by existing structures and systems of classification which were developed for some other use (e.g. search engines, library lists). The latter of these issues is selection and stratification according to externally-imposed, but textual features such as genre, text type, medium, etc. These two issues are likely to overlap somewhat: those distinctions useful for one research question may prove ambiguous for another.

A secondary, yet related, issue is that of metadata completeness: often, texts are sourced from very different places, and come with differing levels of metadata. Normalising and homogenising these metadata is particularly challenging, and may include a loss of resolution.

The temporal aspects of texts are a special case of this. Diachronic corpora seek to represent a 'slice' through time of language use, and this requires a representative distribution of text age across the stated population. Much of this distribution is defined by the production/consumption balance chosen for the corpus: for a given population, how old are texts used on a day-to-day basis? When sampling online, or sampling from historical texts, merely identifying the age of a text is also challenging.

These questions are also important when analysing historical or monitor corpora, where they must be answered in order to allow accurate sub-sampling. Corpora such as EEBO[26] include

such wide temporal coverage that their use often requires identification of a historical context.

Practical Issues

Copyright is one major problem with sampling large volumes of text from any source, especially those already available for-profit from large publishers, who are acutely aware that their entire business model is based on controlling access to their intellectual property.

This is stated in the original documentation for the Brown corpus as one reason for their choice of sampling unit, and often causes 'black spots' in the sampling frame of a corpus, where certain publishers are known for their unwillingness to offer material[126].

As a large number of documents are designed as ephemera, these are often not to be found in archives. As a result, out-of-date ephemera are difficult to obtain retrospectively. Many digital ephemera are targeted at individuals, introducing ethical issues analogous to covert research.

For speech data, and written data that is not intended for public dissemination, privacy must be considered as a legal and ethical issue. Any auxiliary data is unlikely to suggest proportions for documents used within a private context (e.g. notes left on a fridge), or those controlled by formal social structures (such as documents used in many offices).

Some topics and contexts that ought to be represented are difficult to sample due to the cultural taboo surrounding them. This is especially relevant for speech corpora, since speech is often used for informal transactions, and because it is difficult to separate the sampling procedure from the speaker himself.

These issues are particularly pertinent to verbatim recordings in natural settings, such as those used in many spoken corpora. Such recordings, even if taken in public places, may be subject to laws such as the Human Rights Act 1998 which, through guarantees of privacy, may restrict the release of such data to those not present at the time.

Such ethical pressures pose a challenge to unbiased sampling, and one that cannot easily be worked around. The issues surrounding covert research are discussed in more detail in Chapter 4, Section 4.6.6.

Finally, though this is increasingly simple as the world moves towards digital storage, transcription and physical acquisition of data is often slow, difficult, and (thus) expensive.

Paper documents must be scanned, digitised, manually corrected and converted into a normal form. Speech must be gathered in an ethically defensible manner, and metadata about the speakers must be gathered. Electronic documents require significant format conversion.

These problems are easing: digitisation technology, especially that used for paper documents, has recently progressed significantly due to efforts to preserve historical documents, and libraries' attempts to digitise older publications. In some cases, scanners are capable of scanning entire books without intervention. Such advances are not without their own challenges, and significant further work is needed to reduce error rates in line with human transcription[89, 2, 119].

Stratification

Though a problem for categorisation and taxonomy selection, the 'fuzzy edges' of genre, medium, popularity and other covariates often leads to ambiguity surrounding which category a text should belong to.

Some classification schemes apply multiple labels to a text in order to avoid this[161]. This approach may provide a way of producing a more accurate overall classification, but complicates many analyses, particularly quantitative ones relying on regression.

Biber and others[115, 22] recommend that sampling should occur in an iterative process, with the contents of a corpus being used as evidence to weight selection of strata from the next version.

As Varadi[179, 180] notes, this simply doesn't happen. Reasons for this may include the shift in classification priorities, and the time required to re-code and align a previous corpus' annotation structure to that which is to be built.

Auxiliary data from social surveys may be used as a rough indicator for this, though in reality no data source exists that can describe, in social terms, the 'types' of language used (w.r.t. the primary dimensions of sampling for corpora). Questionnaires, ethnography, and/or direct sampling may be the only ways to establish a ground truth for this, but for now it remains an open research question.

Randomness

The pragmatic issues surrounding corpora do not apply equally to all genres, media, social demographics, or settings. This results in the need to apply large qualitative corrections to the sample design, which restricts the ability to use random sample designs.

A prime example of this is the proportions of written and spoken texts in many corpora, and even the proportion of elicited spoken vs. 'natural' spoken texts, due to the difficulty in obtaining consent. Another example may be seen in the BNC's proportion of academic or newspaper texts, both of which comprise a very large proportion of the corpus, yet are read by a relatively small proportion of the population.

In many cases, such as web crawling, nonrandom sampling strategies are used due to the lack of an authoritative (or reliable) central index. In others, such strategies may be the result of other practical issues, such as the location of researchers or legal concerns surrounding certain contexts.

It's possible to augment and correct for a lot of the problems that are introduced through nonrandom sampling, for example, Schafer and Bildhauer do this in their web-scraping corpus building tools, which attempt to stratify their samples by top-level-domain *after* selection[154].

Nonrandom sampling is an unavoidable truth of corpus building in almost all contexts, and with care should not unduly influence the result of a corpus building effort: after all, most fields face similar practical issues.

Sample size calculations are a complex topic, particularly where the variance for a given population is difficult to determine. Due to limitations in power analysis methods for mixed-methods designs, many of the sample size judgements required during corpus construction will

necessarily include some expert opinion.

LNRE models offer possibly the most advanced generative method for this[8, 47], and can be seen as a progression upon Biber's variation analyses of the early nineties. The issues with using such models to determine corpus variability sit neatly in two categories:

1. Selection of interesting variables to measure sufficiency (this, in reality, would be part of the corpus documentation: we can say it is sufficiently representative for frequency counts, grammatical analyses up to n tokens, etc. with little risk)

2. Decisions on how much data is enough data. This is fairly easy to approximate with a sufficiently accurate language model, and even simple binomial approximations prove useful in judging the likely frequency of features.

In spite of the practicality of the techniques surrounding measures of internal feature saturation, few currently do such analyses. This is perhaps because the linguistic researcher himself must usually perform the analyses in terms of his own study criteria, rather than the corpus builder.

The problem of quantifying variation was tackled by Gries[64], where he presents a method for estimating confidence intervals for small-scale features, and uses these to explore the homogeneity of corpora and their internal distinctions. Therein he focuses not on simple frequency variation, but on grammatical properties that, he claims, are more representative of the sort of inquiry made using corpora. He starts by using the $Z - score$ centrality measure, noting the limitation of existing partitioning schemes [64, p.123]:

> *In order to measure the homogeneity of a corpus with respect to some parameter, it is of course necessary to have divided the corpus into parts whose similarity to each other is assessed.*

He then goes on to augment this method by way of bootstrapping, in order to identify regions of maximum homogeneity within a corpus in a ground-up fashion. Such permutation testing offers a ground-up mechanism for identifying genres, at the potential cost of interpretability.

Sampling / Analysis Unit Mismatch

Any sample is, strictly speaking, best analysed in terms of its sampling unit. Any disagreement leads to a reduction in how free each data point within the sample is to vary[94, 117, 136].

Biber's initial assessments of language variation in 1988[20] examined the suitability of the 2,000-word sampling unit by splitting each sample and comparing the relative frequency of features in each half. He determined that, if they were the same, then the sampling unit size was sufficient to represent a given feature. This lead to the conclusion that corpora were adequate for inspection of 'small-scale grammatical features', something that seems to be borne out by the success of computational models in this area (POS tagging, etc.).

This issue is often phrased in terms of dispersion[94, 88], i.e. the tendency for a feature to be represented evenly in all documents throughout the corpus. Dispersion is something that manual inspection of concordances and corpus data renders particularly transparent, as it is

often possible to see that all instances of a particular idiom are traceable to a given author, or all coverage of a certain topic is from one publication. This is a prime example of the unit of analysis being far smaller than the sampling unit, a problem that is receiving increasing recognition.

Evert, with his library metaphor[46], describes a sampling policy for corpus linguistics that avoids this disparity. In it, he describes randomly sampling progressively smaller units in a virtual library, moving from books through pages to sentences.

Since many statistical problems arise from the disparity between sampling and analysis units, a particularly problematic instance of this issue is the use of bag-of-words (BOW) models. These model language without taking into account order, and as such would be best applied to a corpus sampled at the word level: any section of text beyond this is going to exhibit 'clumpiness' effects that are beyond the comprehension of the model.

The prevalence of BOW models is such that many people choose to phrase their objections in terms of the accuracy of binomial models of language frequency[94, 45, 47]. Undoubtedly they have a point—more complex LNRE models are capable of far more useful inferences, however, it's worth noting that only a model that can integrate the linguistic influence of word 1 upon word 2,000 in a sample will truly justify a 2,000-word sample[4].

One part of the problem is caused by a fixation on word frequencies. A given corpus, 100,000 words in length, may comprise just 50 texts. For the purposes of many analyses involving person-person variation, it can be said that we really only have 50 data points. Though care is often taken to sample these texts broadly, the idea of targeting a corpus in terms of its word length, rather than the number of samples, leads to extremely poor suitability for analysis of many features.

The problem of quite what sample size to choose is a trade-off: if we were to sample single sentences for our 100,000 word corpus, we would soon require thousands of samples, each from a randomly selected and carefully-stratified source. If we wished to perform a complex analysis of narrative structure within those sentences, we'd find there is insufficient data.

Except in certain cases, there is a plateau of difficulty for sample size: sampling 2,000 words from a book incurs very similar levels of practical obstacle than does sampling 100. Sampling units should be chosen in accordance with the complexity required by researchers of the time—for example, those working in information retrieval and NLP fields will demand relatively large sample units by comparison to many researchers in linguistics.

Corpora should, ideally, be defined in terms of the number of *datapoints*, rather than words. This is one area where compatibility with existing corpora and techniques is arguably damaging to the final result, and where documentation should be clearer in guiding valid use.

2.2.6 Validity Concerns in Corpus Analysis

One of the primary reasons for using quantitative methods in research is their objectivity and empirical basis. This is especially the case in linguistics, which seeks to generalise about a social

[4]This is a quantitative form of Hoey's argument that whole texts are necessary to truly describe human expectations of word use[75].

property that is difficult to quantify or relate to other users.

In addition to the above procedural concerns, manual inspection and summarisation of corpora (or collections of features extracted by corpora) often leads to situations where errors of human judgement may imply certain findings. These cognitive biases are widely recognised in many cases, and have been identified in other fields as common causes of error[84, 116].

Many of these biases are fairly minor, and scientific methods are designed to counter-act their effect. Nonetheless, qualitative analysis, poor corpus design, and presentation of certain features once extracted for inspection, can introduce their effects.

- **Insensitivity to Sample Size**
  People are liable to underestimate variation in small samples. Manual inspection of particularly intricate features extracted due to their grammatical form, especially from subsamples of the corpus (such as the spoken part) are likely to qualitatively imply false results[143].

- **Clustering Illusion**
  A tendency to find patterns where they are none (aphophenia). This is more of an effect when seeing larger data sets, such as when inspecting frequency lists or concordances.

- **Prosecutor's Fallacy**
  Assuming conditional probabilities indicate unconditional ones, and vice-versa. This might be less prevalent, but when conditioning on a certain feature, subsample of the corpus, etc. it is common to assume that the trends identified are indicative (or anti-indicative) of similar trends in the rest of the corpus.

- **Texas Sharpshooter Fallacy (post-hoc theorising)**
  Testing hypotheses on the data they were derived from, i.e. inspection then derivation then testing. Also called 'data dredging', this is a risk of large-scale community reuse of corpora. A solution to this is to reuse sample *designs*, but not the actual data itself.

- **Availability Heuristic**
  The tendency for people to mentally overestimate the probability of features which they can immediately recall examples for[176, 154]. This effect is likely to occur when examining corpora qualitatively, particularly if features are inspected in response to searches on particular features.

Many of these concerns are difficult to address without fully quantifying or automating analysis stages that are currently performed manually, something that is often a technical challenge. Though reduction of their incidence is the goal of a corpus builder, there is often little that can be done directly prior to analysis.

The quantitative stages of corpus analysis are also not free from statistical challenge. Many of these issues surround the use of models making flawed assumptions about variability[94], something that is a direct result of disagreements between sampling unit and analysis unit.

One approach to this is advocated by Wallis[181], who proposes using a baseline linguistic feature as a control. This serves to normalise the frequency against which one is comparing

during statistical tests, but remains subject to issues of low power due to use of samples that contain too few independent sources.

Use of better-informed linguistic models in order to take account of non-orthogonal variation in linguistic features is also possible. One method of doing this is to adjust for (or at least recognise qualitatively) the dispersion of features across samples[65, 66]. Research is needed in order to best use these measures for qualitative interpretation (dispersion is already displayed in some popular tools[4, 147]) and to adjust existing statistical procedures.

2.3 New Technologies

2.3.1 Web Corpora

The rise in popularity of the web introduced a significant new source of text for linguists. Whilst computerised text has been included in many conventional corpora (in the form of email and other direct communications), the ubiquity of the web offers a source of digital-format documents that now cover many subjects and genres.

Due to widespread and diverse use of the web, such corpora span many populations—some efforts focus on describing the web itself, and representing the population of documents online as its own population. Others are able to focus on more general representation, or offer metadata sufficient to retrieve special-purpose corpora. It is certainly the case that any modern corpus claiming to be representative of production or reception should include web data.

The first efforts to construct corpora from web data[96] focused on the production of tools and resources to download, clean, and index web data at large enough scales to be used in general purpose corpora. Many of these problems have now been solved to some degree, with the WaC movement producing a number of web-specific corpus tools:

- **Crawlers**
  WaC-specific crawlers have been produced that are able to control their behaviour according to linguistic properties[154], and with the intent to provide an overview of proportions online.

- **Boilerplate removal tools**
  Tools such as JusTexT[139] and BoilerPipe[103] are able to remove the 'boilerplate' of websites in order to leave just the content areas.

- **Genre classification schemes**
  Taxonomies have been produced that include the new genres found online, or those new forms of text such as blogs. Some of this work has come from search engine designers, eager to improve their results[131], as well as from linguistics proper[160, 153, 161].

- **Retrieval tools**
  Tools such as BootCaT[14] and search engines such as WebCorp[148] allow access to increasingly large-scale web resources.

The ease with which users may access web data is of particular value: this has allowed for corpora of unprecedented scale such as WaCky[15] and COW[155] (which contain tens of billions of words), as well as for near-instant corpus building and replication.

Many WaC efforts focus on sampling the web itself, in order to represent users' experience thereof. This approach leads to sample designs that focus on uniform coverage of top-level domains, or particular types of web resource, in a manner similar to that of search engine crawlers.

Performing this task in a robust manner is particularly challenging due to the lack of any central index for websites at this level. As such, it is common to oversample and then select documents after-the-fact using scoring techniques[156, 154]. This remains challenging due to the difficulty in obtaining consistent metadata for the web.

A contrasting approach is that of using existing indices to retrieve data, usually by using commercial search engines. This is the approach taken by the BootCaT tool, which relies on repeatedly requesting URLs from the Microsoft Bing search engine conforming to a number of seed terms.

This could be seen either as an attempt to construct a corpus in the image of the seed (hence the 'bootstrapping' in the name), or, under conditions of sufficiently general input, one to represent the web itself, using the search engine as a 'lens' through which to view web documents. This is a distinction drawn by the source of variation: the seed terms themselves, or the web's ability to return the desired data. The primary purpose of this was to find a ready source of documents, and as such conventional sample designs were still used as gold standards.

Performance of such methods was examined by Kilgarriff and Grefenstette [97, p. 343], stating regarding representativeness:

> *The Web is not representative of anything else. But neither are other corpora, in any well-understood sense.*

This is something I will revisit in Chapters 3 and 6, as WaC methods are examined in more detail.

2.3.2 Intellectual Property

The legality of releasing large volumes of text has always been an issue for corpus builders, however, much of the time this issue has been handled by an intermediary: the publisher, able to provide rights to many documents at once. WaC methods access content by many publishers, often where rights are poorly-stated or belong to owners which cannot be contacted (especially in bulk). This issue is complicated further by the international nature of the web, and the style of hyperlinking and content embedding[72].

One approach to this is to limit access to end users, providing an interface which still permits some level of inquiry whilst restricting large-scale access to the whole volume of data. This is the approach taken by BE06[9], which is available through CQPWeb[71] even though it contains data that is nominally copyright to other parties.

A second approach is to provide lists of URLs from which the original data may be retrieved. This is less legally troublesome than that taken by BE06, yet also requires a significant amount of work on behalf of any replication.

Finally, corpora such as the Google Books corpus[60], COCA[37] and the COW corpora[155] are offered in a jumbled form, designed to retain frequency information without the ability to recover full texts. The corpora mentioned here are particularly large, implying that they are unlikely to be used by researchers reading the text directly due to the large number of results for even very specific queries.

I consider that these approaches miss the point somewhat, as the ease of retrieval WaC brings make it possible to perform full scientific replication: retrieving new data conforming to the same sampling policy, rather than repeating the data verbatim.

2.3.3 Sample Design

The web, and its ease-of-access, also enables new sample designs.

The most obvious of these is the automation of monitor corpus maintenance—this is analogous to the problem of maintaining search engine indices, and as such there is significant literature from that area focused on the scheduling revisits to pages and ensuring coverage across domains. These techniques are now used in lexicography to keep dictionary resources up-to-date[130].

The availability of translated documents online also provides easy sources of data for constructing parallel corpora, a task that can be almost completely automated due to web page markup[149].

The ability to automate retrieval also means that diachronic designs can be automated to some extent. This is explored and developed further in Chapter 3.

2.3.4 Documents of Digital Origin

With the rise of the paperless office[5], even documents typically accessed in physical form are authored and stored digitally. This has now extended to almost every form of textual information.

This has two main advantages: firstly there is greater coverage of conventional formats such as books. Secondly there are entirely new opportunities to sample and meaningfully process things that have never been available before, like flyers, video with overlays, etc.

The realisation of increasing quantities of data as native-digital objects means that it is possible to take copies with relatively low costs, and without depriving the original owner of the resource for any great length of time.

The increased breadth of use also lends itself to multi-modal corpus designs, effectively expanding the coverage of a corpus to better suit its population.

[5]For an illustration of how well this design principle worked, refer to the contents of your own desk.

2.3.5 Life-Logging

Life-logging is an activity that is focused around gathering, organising, and using a continuous record of the data encountered in everyday life. It has been developed with two main focuses, both of which may lend value to the process of corpus building and sampling:

- Entertainment—Many people have, since the mid 1990's, broadcasted significant portions of their life online, something that has risen in popularity to the point of spawning consumerised applications for the purpose (Justin.tv, uShare). Methods used focus on audio-visual broadcasting, as the output is matched closely in format to reality TV.
- Information categorisation and extraction for personal use—this has been the main focus of the academic community (and, in one notable case, DARPA[3]), and has spawned many projects that focus on digesting and operationalising lifelogging data. Typically such efforts are less focused on audio-visual data, since it is prohibitively difficult to process.

With the availability of powerful portable devices such as tablets and phones (and wearables such as Google Glass), life-logging techniques that have conventionally been restricted to only a few individuals worldwide due to technical requirements or practical limitations are becoming increasingly viable as sources of information for many people.

These techniques offer an approach to sampling that, whilst explored by many other fields, is typically seen as expensive and involved. One's own logged history may be a useful source of data for systems that interact using NLP techniques (in order to mimic one's dialect and idiomatic language use more closely), or the language proportions of social groups may be more accurately determined for scientific study. The value of such personalisation is already proven in many contexts with more limited interaction methods, for example speech recognition (personalised phonetic models) and web search (Google and others' personalised results). Further to the benefits of being able to gather real data more easily, life-logging allows us to peer further into social contexts with less disruption, yielding higher quality data.

Where language metadata is needed (rather than verbatim text), many life-logging technologies support discarding of any identifiable information on-the-fly—there are techniques for storing only irreversibly scrambled audio such that the characteristics of speakers may still be identified [112], or devices with sufficient power can simply store summaries of the events they observe, discarding the data itself. The decrease in ethical sensitivity associated with such measures further reduces the boundaries to wider sampling of a population, something that may be used to improve and adapt existing languages resources.

Such life-logs, even heavily anonymised, offer new opportunities for balancing corpus strata, as well as providing a perspective on use that is more richly annotated with contextual metadata. These methods are reviewed in more detail in Chapter 4.

2.3.6 New Sampling Opportunities

Miniaturisation of technology, and increases in computing power, gradually unveils new opportunities to sample data. At one level, this may make possible voice recognition and

transcription of multiple subjects, or analysis of data in-place, without transcription. At another, computationally-expensive techniques such as bootstrapping become increasingly applicable to samples large enough to be used for linguistic purposes.

Access to Technology

The ubiquity of technology makes accessing a diverse population of language users with little concern for geographic limitations. Submissions of textual data may quickly be acquired via the internet, and populations of people otherwise unrepresented in corpora may be sampled this way.

This has also had other effects, such as the tendency for a single conversation online to include speakers from many countries, cultures and demographics (often without even being aware of that fact).

One effect of this has been the ability to recruit study participants from across the world at relatively little cost through the use of crowdsourcing platforms such as Amazon AMT[83] or Crowdflower[50]. This widespread ability brings with it a significant number of analytical challenges, many of which are still being worked upon, and many of which, such as the problem of managing inter-annotator agreement with many participants, apply nicely to the problem of sampling quality corpus data in a distributed manner.

Indexing and Access

The improved power and utilisation of large-scale computing machinery opens up possibilities for more complex examination and extraction of data from existing lists. A prime example of this is the pooling of data in and around search engines such as Google, which have become a source all of its own for many researchers[95].

Once data is acquired, data warehousing techniques open possibilities for examination using many more covariates than has previously been possible, allowing very complex research questions to construct meaningful subcorpora with relatively little effort or time overhead.

This bulk analysis of heterogeneous data is regarded as a growth area in many circles, though seems to have been under-utilised in academia due to its low quality compared to traditional scientific samples, and the low replicability that implies. Whether 'big enough' big data is ever useful in a scientific context remains to be seen, however, this is arguably something WaC already embraces.

2.4 Summary

Conventional corpus sampling techniques are, fundamentally, based on those used elsewhere. There is, however, much confusion on the topic, and unwarranted debate over whether or not corpora are comparable to typical samples.

Much of this debate has stemmed from tradition, being sparked by the complexity of corpus designs such as Brown's, which relied heavily on expert opinion and inclusion of text according

to a deliberately non-representative policy. This subjectivity is, in reality, a reaction to practical challenges that faced the first corpus construction efforts.

Many of these practical challenges have either changed in nature, or been solved. Digitisation and POS tagging are now largely automatic affairs, and documents selected for inclusion are often born digital. For UK researchers, there is even relief from a number of intellectual property issues.

Unfortunately, however, some have proven harder to solve. There are still large challenges to the problems of population definition, and definition of a widely-agreed-upon taxonomy for genre definition. Additionally, access to documents, even where technically possible, is often restricted by private corporations, or filtered through systems such as search engines that act as black boxes.

These remaining challenges are still limiting the corpus building process, and prevent the use of the most quantitative random sampling techniques. Even those that are part expert-informed, such as stratified sampling or cluster sampling, require widespread agreement on taxonomy selection. It is unlikely that there will be any one answer to such theoretical issues, as these must be defined relative to the research question.

The Web-as-Corpus movement offers a democratisation of the corpus building process by easing a number of the retrieval and enumeration problems[6]. These come with a significant challenge of their own, however, in that each user must then perform all of the stages of a corpus construction effort: from unambiguously specifying population parameters to cleaning and preparing the data.

This process can, at least, be automated. In so doing, it is possible for researchers to vary their corpus designs and tailor them to meet individual needs. For this to occur, tools and methods must be produced that are able to take design goals and, without significant further effort, produce corpora from them.

The role of this thesis is to explore this space, investigating corpus sampling and identifying areas where WaC approaches may mitigate remaining issues. The ultimate goal is to provide methods and tools which allow users to rationalise and operationalise design criteria to produce their own valid and useful corpora.

[6]One that is, as lamented by Pedersen, rarely realised in practice[137].

Chapter 3

Longitudinal Web Sampling

The open publishing model of the web leads to a far faster turnover of documents compared to most conventional printed sources. In addition, the lifespan of a document is controlled by its author[7], rather than by any publishing party, often requiring the continuous provision of funds merely to stay available. The natural conclusion of this is that documents available online are there in a transient, time-limited sense only—in order to uniquely identify a document online, both a URL and time are needed.

The phenomenon of documents becoming unavailable over time, known as 'link rot', has been studied at length by those who seek to index the web, most notably search engine designers. Lately, work has also been done to identify flaws in the longevity of scholarly collections[99].

Document change over time also has important ramifications for those working with web data. Aside from changes in ownership of domain names and site restructuring, many web pages undergo gradual revision and updates, particularly those in the news genres. Again, whilst there has been study of these events for the purposes of crawler design, the linguistic impact is less well known.

As WaC methods become more widely adopted, and other general-purpose corpora seek to include web data, the distribution of documents through time becomes an important sampling issue: the age of information when read, and the longevity of it when written, both affect the veracity and relevance of any conclusions. This is of particular importance if using web data to match another sample design, as such effects must be controlled for at a later date. There is also the possibility that this attrition correlates with genre, meaning that bias creeps into the dataset over time.

Current crawler design (and thus the contents of search engine indices) focuses on keeping the most current page at any given time. This may be insufficient for many scientific purposes, which must know the context and contents of each page over time, and do so in a manner comparable to other documents from that time period.

This chapter starts in Section 3.1 with an examination of current open-source corpora, and their likely half-lives online. This is a motivating example to frame the introduction of the

[7] Though this situation is quickly changing as people increasingly use blogging platforms and social media to publish content.

LWAC tool (Section 3.2), which provides a mechanism for longitudinal sampling of large corpora over time. The suitability of this tool for long-term use is evaluated therein, before a summary (Section 3.3).

3.1 Document Attrition

Those using corpus data are increasingly turning to the web as a source of language data. This is unsurprising given the vast quantities of downloadable data that are readily available online. The Web as Corpus (WaC) paradigm[97] has become popular for compilers of corpora for lexicographic use, replication of standard reference corpora as well as for studies of specific online varieties of language.

Two general models of WaC have emerged. Using the first model, 'browsing', corpus compilers collect data from a set of given domains and select whole or part texts online and incorporate them into their corpora (e.g. the BE06[9] corpus comprising material published in paper form but found on the web).

The second model, 'searching', sees the web through the lens of search engines, and is typically used to compile domain-specific corpora from a set of seed terms which are used to locate web pages for incorporation into a corpus (e.g. the BootCat & WebBootCat tools[14]). In some cases, both approaches are combined, using searching for general language seed terms to produce reference corpora[98].

Collecting corpora online raises legal questions regarding redistribution rights. Consequently, many compilers choose to make data available only through a web interface (with restricted access for fair-use) or by distributing URL lists (known as *open-source* corpora[159]).

Online content changes much faster due to the decentralised, open publishing model of the web, which may have a serious impact on two aspects of the WaC paradigm: availability and replicability.

As websites change, the URL lists need updating to reflect new locations. Worse, websites or pages may be completely removed, thus the corresponding part of a corpus is no longer available. This attrition of documents through time affects both users of open-source corpora and those attempting to interpret the results of studies using corpora that were built just a few years ago.

Though many studies have looked at the life-cycle of web pages in general, these typically focus on the integrity of websites or specific repositories of information, rather than the documents and the language contained within.

Koehler[100, 101], through four years of weekly sampling, found that just 66% of their original seed URLs remain online after a year, with this proportion dropping to around 31% by the end of the study. Koehler started his study in the early days of the web (December 1996) using a relatively small sample of only 361 URLs. His analysis found statistically significant differences in the type of page as well as variation across top-level domains (TLD).

Nelson & Allen[134] found that only 3% of documents in digital libraries become unavailable in just over a year. This is perhaps unsurprising given the aim of such projects but serves to

illustrate the degree of heterogeneity between types of document host.

Bar-Ilan and Peritz[10] present a slightly different method of study, searching for the same information where URLs became unavailable. Though they primarily investigate the growth of the web in their analysis of the availability of information, they find a similar amount of naïve URL attrition to the above studies after a year.

Studies involving academic paper availability mirror open source corpora in that they use references in favour of original text, however, the centralised administration of academic repositories is notably in contrast to most web resources. Nevertheless, there has been much work into this area, some of which is comparatively reviewed in a paper by Sanderson et. al.[152] These studies, spanning years from 1993 to 2008, illustrate that even institutions charged with keeping an accessible record of information are still subject to rates of attrition in the region of 25–45% over five years. The reasons for this loss are not investigated.

Despite these enquiries, very little work has been carried out to estimate the effects that document attrition has upon corpus content. Sharoff[158] touches upon this in his WaC work, presenting a preliminary analysis of attrition within corpora generated by searching for 4-tuples. Although his studies lasted from one to five months, and contained modest sample sizes (1000), they indicate a rate of loss that is below that of other studies (just 20% per year in the month-long study), suggesting that the selection of documents may have significant effects upon document attrition rates.

Rather than analysing web resources in isolation of their linguistic uses, I outline a preliminary analysis of what I term 'document attrition' relative to a number of corpora of differing construction. This is done in a manner similar to those using open-source corpora: the retrieval of URLs, rather than information.

3.1.1 Methods & Data

In order to measure document attrition across a number of linguistic sources a set of corpora were selected, chosen due to their differing constructions and ages, and downloaded using a process that closely approximates an end user's view of the web. Statistics on the availability of these documents were then annotated with a series of URL-related variables for analysis.

Data Summary

Data were taken from four open-source corpora (outlined in Table 3.1), each of which consist of a sample of URLs referring to web resources. All of these corpora are cross-sectional, representing data from a short period, however, only the BootCaT-based corpora are built using a script that is likely to sample quickly enough to count as a true point sample in the context of this study.

BE06[9] was built as a conventional, hand-selected corpus designed as an update to the LOB[86] and FLOB[77] corpora[8]. It contains texts from sources published in 2006 but also available online.

[8]The author wishes to thank Paul Baker for providing the URL list for this study

Corpus	Date	Size (URLs)	Sample Period	Construction
BE06	2006	473	1 year	Browsing
Delicious	Sep. 2009	630,476	1 month	Browsing
Sharoff	2006	82,257	hours	Searching
BootCaT	Sep. 2011	177,145	hours	Searching

Table 3.1: An overview of the corpora selected for study.

The Delicious corpus represents a sample of links posted to the front page of delicious.com[9] during the whole of September 2009.

Sharoff's corpus is the same one used in his 2006 paper on open-source corpora, and is built using modified BootCaT scripts from a series of 4-tuple seed terms selected from the British National Corpus. As with BE06, the month of construction was assumed to be June (half-way through the year), in order to minimise potential error.

The BootCaT corpus is built for this study from 4-tuples that are themselves built from the same terms as Sharoff uses for his 2006 study, using updated versions of his original scripts[10].

Downloading Process

The process of recording the document's status was relatively simple: a small piece of custom software was written to download documents from an open-source corpus at regular intervals. This tool was configured to mimic requests made by common web browsing software in order to emulate a typical user's visit to the document. Handling SSL, cookies, and referrer links in a similar manner to a user following a bookmark allows us to assess more accurately the content, avoiding tricks that exploit search engine crawlers and other bots.

The tool used here was later developed into a more complete solution for investigating temporal effects on the web, and this is discussed in more detail in Section 3.2.

Taking this user perspective, the notion of document availability becomes slightly more complex. Redirect requests were followed up to a depth of five, as recommended in the HTTP specification and commonly implemented in browsers. Since I do not account for content changing in this experiment, failure was taken as receiving a final HTTP status code other than 200, or a network timeout (60 seconds was the timeout used for DNS and TCP)[11].

Both metadata and original response details are stored by the download software. This experiment will focus on features of URLs, such as the presence of GET arguments[12] in the URL string, meaning that the resource is likely to be dynamic. These features have been chosen to indicate aspects of web hosting and affiliation that are likely to vary between users with both different reasons for uploading their content, and different degrees of technical expertise.

[9] A social bookmarking site where users post and exchange links.
[10] As retrieved from http://corpus.leeds.ac.uk/internet.html
[11] Timeout errors also occur stochastically due to routing policies, and are impossible to avoid entirely when downloading resources in bulk. The download process was tuned to minimise this source of error.
[12] Parameters appended to a URL string, typically used to control dynamic scripts

3.1.2 Results

Table 3.2 describes the availability of documents within the corpora, as sampled on the 21st October 2011. This forms two data points, the former representing no attrition when the corpus was first compiled. Document lifetime statistics are calculated assuming exponential decay: both the half-life ($t_{1/2}$, the time it takes for only half of the original corpus to remain available) and the mean lifetime, τ, are provided.

Corpus	Age (yr.)	Loss	$t_{1/2}$ (yr.)	τ (yr.)
BE06	5.3	42%	6.5	15.8
Delicious	2.1	7%	17.8	42.4
Sharoff	5.3	34%	8.6	16.4
BootCaT	0.08	0.8%	4.8	20.4

Table 3.2: Loss from corpus inception to October 21st, 2011.

The large differences in corpus half-life are revealing—the Delicious corpus has significantly lower loss than the others. This is ostensibly owing to its construction: users are likely to bookmark resources that are useful (and hence are well-established, popular sites), in contrast to BootCaT's uncritical selection or the deliberate *document*-seeking (rather than *information*-seeking) represented by BE06.

The difference in half-life between Sharoff's corpus and my own BootCaT-derived one is harder to explain. Though both are large corpora built using similar methods, they were sampled years apart with heavy influence from search engines which, will have updated significantly in this period. It is also possible that 31 days is an insufficient period to achieve an estimate for attrition that is representative of a full year's loss, implying that a pattern may be evident due to external influences (such as hosting renewals around the commencement of the tax year).

The half-lives of our samples are above those stated in other, more general, studies (Koehler reports a half-life of 2.4 years, for example). This may be the result of bias introduced when deliberately omitting non-document portions of the web, such as navigation pages or images. Another influence is the age of many attrition studies, as it is possible that, with reduction in the price of web hosting, resources simply remain online for longer.

The relatively high rate of attrition in BE06 is surprising given that it only features documents that were already in print, which ostensibly reside in archives or the websites of large institutions (shown to be relatively nonvolatile in other work). One possible counter argument is that BE06's sampling policy was to take documents *published* in 2006, rather than merely being available, such that these samples have witnessed the initial steep descent on the document survival curve.

The diversity of status codes returned varies significantly between corpora, with older ones showing more intricate and descriptive modes of failure (such as code `410 Gone`). Delicious exhibits differences to the other corpora, exhibiting codes that are presented by WebDAV[13] and similar systems unlikely to be crawled by search engines.

Each of the corpora exhibit similar distributions of each top level domain (TLD), though the large differences in sample size make formal comparison difficult. The overall distribution of

[13]WebDAV is used for modifiable content, usually the presentation of revision control system data such as subversion repositories.

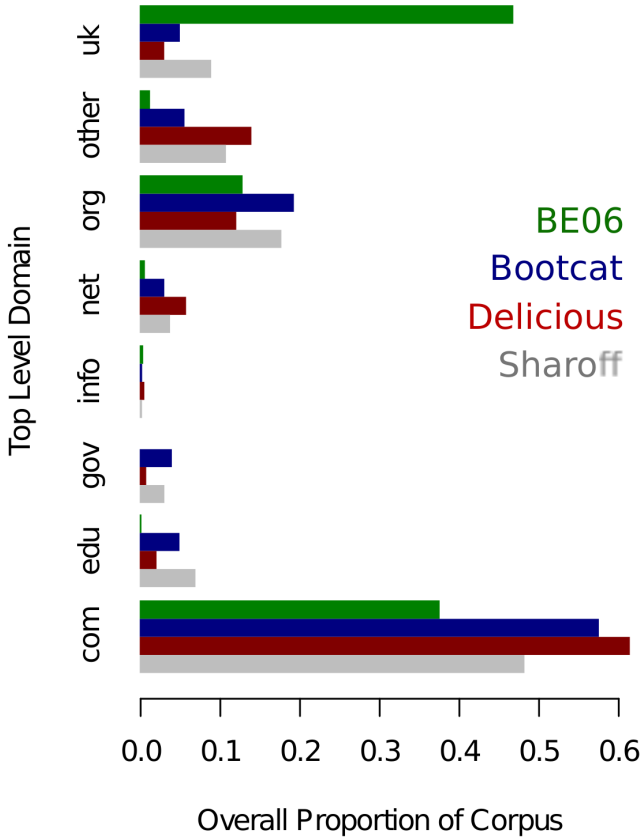


Figure 3.1: Overall distribution of top level domains.

the more popular domains is provided in Figure 3.1. This indicates that `.com` dominates the selection across corpora, with `.org` following. Only BE06 differs from this distribution in its selection of `.uk` addresses, however, it has been deliberately biased this way so as to represent British English.

Other studies have identified statistically significant differences in the rate of attrition between the major TLDs[101]. Though each corpus exhibits a dependence between these TLD groups (chosen to represent the vast majority of each corpus' content) in a χ^2 indepdence test ($\chi^2 > 33.8; p < 0.01$ in all cases), generalised linear models reveal that the nature of this dependence varies greatly between corpora, indicating that this estimate is far too simplistic to represent the real causes of attrition.

The shape of the empirical distribution for path length of the URL is shown in Figure 3.2. Delicious.com users may be expected to bookmark top-level domains with relative frequency, but the difference between the two BootCaT samples is more subtle, perhaps an effect of search engine changes. The preponderance of introductory or 'launch' pages in the Delicious data set may also go some way to explaining the longevity of its content—it seems reasonable to presume that top-level pages remain online for longer (though also perhaps that they change more often).

Taking the presence of GET-arguments in a URL as an indicator of a page being dynamic, a number of effects may be seen across the corpora. The BE06 corpus had 24% of all links

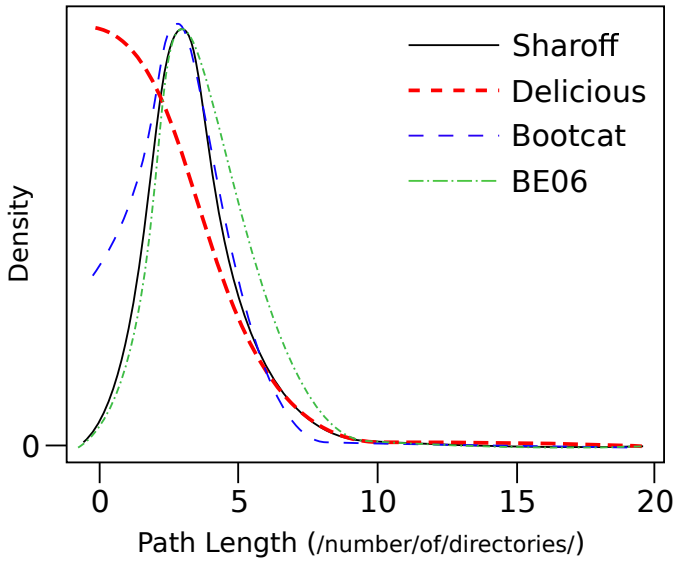


Figure 3.2: Differences in the empirical distribution of URL path length.

dynamic, exceeding Sharoff's at 17% and Delicious & BootCaT at 8% and 6% respectively. This difference is probably due to the selection of published documents, since the compiler was seeking specific materials within sites rather than attempting to retain the location of a resource (as with Delicious) or sampling randomly from URL-space (as with BootCaT). Differences between the two BootCaT-based corpora may reflect changing weights within search engine algorithms.

3.1.3 Discussion

These preliminary results indicate that the process of corpus compilation, by introducing deliberate bias into the content (the selection of full documents, filtering of navigation pages, etc.), impacts the observed document attrition rate. These biases have been evidenced by the URL features alone, raising interesting questions about the effect that collection strategies have upon corpus integrity—should the tendencies of different groups of web publishers be factored into sampling strategies for open-source or subject-specific corpora?

The ramifications of these biases for the WaC availability sampling strategy remain an open question—does 'searching' for links imply a minimum age, and hence a pre-existing skew towards certain content?

There are indications that sampling a cross-section of production, rather than consumption, observes the initial steep decline in document availability that is inherent in most survival distributions. It is possible that these effects are minimised by the WaC approach, and are actually more pronounced in conventional, offline, corpora: search-based sampling may compensate for this effect by weighting reliable and established websites through the algorithms used to rank relevance, though further work is needed to establish the degree to which this occurs. Another possible effect is the disproportionate availability of archived, out of copyright, documents.

In the best case, where all linguistic areas decay at a similar rate, this merely causes us to question the validity of corpora sampled even a short time ago. Where conclusions depend upon

the changing nature of presentation or context this change is likely to be more pronounced, as websites taken down are notionally unlikely to be uploaded without some degree of change.

In order to operationalise the changes observed over time, a more detailed understanding of content changes is required that is able to differentiate between mechanisms of attrition, such as site redesigns, boilerplate changes, and total loss of information. In order to better understand these, a more comprehensive method of sampling is required to inform survival models that are more advanced than simple exponential decay.

3.2 LWAC: Longitudinal WaC Sampling

Many sampling efforts for linguistic data on the web are heavily focused on producing results comparable to conventional corpora. These typically take two forms: those based on URL lists (e.g. from search results, as in BE06[9], BootCaT[14]), and those formed through crawling (e.g. enTenTen[14], UKWaC[49]).

Though initial efforts in web-as-corpus (WaC) focused on the former method many projects are now constructing supercorpora, which may themselves be searched with greater precision than the 'raw' web, in line with Kilgarriff's vision of linguistic search engines[93]. This has led to the proliferation of crawlers such as those used in[155] and WebCorp[15][148].

This approach, with its base in a continually-growing supercorpus, parallels the strategy of a monitor corpus[163], and is applicable to linguistic inquiry concerned with diachronic properties[92].

Repeated sampling by crawling, whilst balanced linguistically, omits subtler technical aspects that govern consumption of data online, most notably the user's impression of its location, as defined by the URL. Low publishing costs online, paired with increasing corporate oversight and reputation management (both personal[120] and professional[122]), have lead to a situation where this content is being revised frequently, often without users even noticing.

The nature of within-URL change have been studied from a technical perspective by those interested in managing network infrastructure, compiling digital libraries[177], and optimising the maintenance of search engine databases[101]. The needs of these parties are quite aside from those of corpus researchers, however, since they focus around a best-effort database of information, rather than a dependable longitudinal sample with known margins for error.

Current methods for sampling language change online include:

- **Crawlers/Revisiting**
  Continuously crawl pages, adding new ones as links are updated and revisiting old ones according to heuristics;

- **Diachronic corpora**
  Build two separate corpora using the same sample design at different points in time;

- **Monitor corpora**
  Continuously add new material to a corpus as it is created, agnostic of change;

[14] http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen
[15] http://www.webcorp.org.uk/live/

- **Subsampling supercorpora**
  Build a large corpus and select documents from it according to a bimodal distribution of document age;

- **Feed corpora**
  Crawl pages on publication date (a form of monitor corpus).

Each of these methods is subject to a number of downsides, making them difficult to use for many scientific purposes, or requiring significant resources and forward-planning (thus putting them out of the hands of most researchers). Some issues, such as irregular revisiting of pages and the lack of detail stored about network-level metadata, complicate the process of analysis. Most corpus formats are also not annotated by time, meaning that analysis often requires repeated export and comparison—tools for which must be produced on a case-by-case basis.

I present here a tool, LWAC, for this form of longitudinal sampling, designed to maximise the comparability of documents downloaded in each sample in terms of their URL rather than content. To accomplish this, a cohort sampling strategy is used, as illustrated in Figure 3.3, to get full coverage over a list of URLs at the expense of sampling new content.

This approach allows for reliable re-visits to each member of the sample, and thus the construction of vertically comparable data points, whilst making short-term effects visible by revisiting each link. Such a sample design should repeat each individual sample as quickly as possible, so as to minimise the time differences between documents within.

This allow a user to investigate how language may change in relation to technical and social events in a way that mimics the experience of many end users, and offers a useful perspective on many epistemic problems of WaC methods, to determine:

- The portions of web pages that typically change as main content regions;
- The impact of social feedback and user generated content on page content;
- How censorship, redaction and revision affect website contents;
- Website resource persistence and its relation to linguistic content;
- How institutions' publishing policies affect reporting of current events.

LWAC is designed to construct longitudinal samples from URL lists, using only commodity hardware. It is designed with 'full storage' in mind, that is, recording everything about each HTTP session in such a way that it may later be exported and accessed in a parsimonious manner.

LWAC is based around a central corpus store that is indexed both by time and URL. Each sample consists of all datapoints in the corpus, and begins at a specified time: this time is aligned to a sampling interval.

Figure 3.3 shows the logical structure of the corpus: the blue bars represent each sample. The primary technical challenge lies in maximising the parallelism of each sample, thus improving the comparability of datapoints therein and allowing for a smaller sampling interval. Samples that overlap their sampling interval will not be stopped: instead the next sample will be delayed until the next 'round' interval. This eases analysis by ensuring the integrity and temporal alignment of each sample.

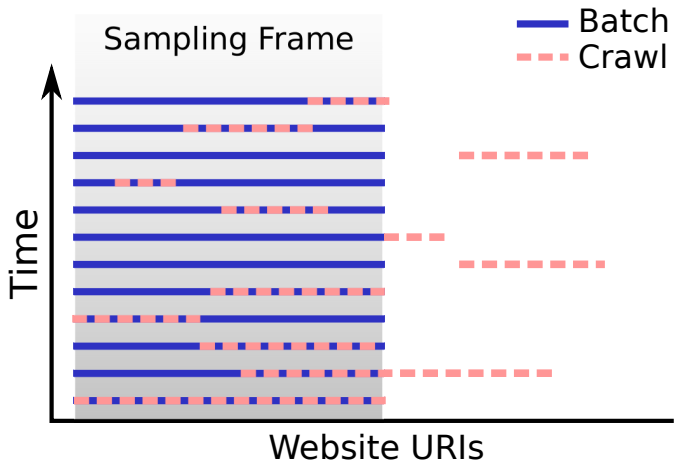


Figure 3.3: URL coverage for batch and crawl

Documents are downloaded in a random order, so as to avoid systematic variation in sample times across links.

3.2.1 Design & Implementation

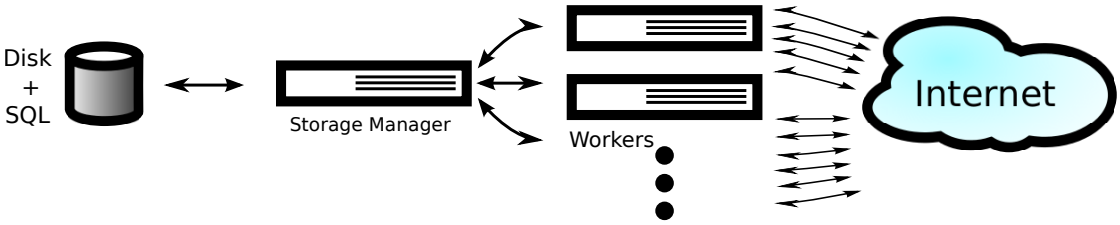


Figure 3.4: System Architecture

In order to perform the parallel retrieval stage quickly, LWAC is constructed as a distributed system, with a series of workers performing the page retrieval and passing batches of data to an offline system for further processing. This architecture, shown in Figure 3.4 is extensible until the limit of contention for the storage manager's network connection is reached, which in practice is encountered at many tens of worker nodes. In addition to the use of multiple workers, each worker performs many parallel requests[16].

The storage manager is responsible for all scheduling and data integrity: workers are treated as (relatively) untrusted actors, and allocated batches of links on request. The scheduling system monitors performance of each worker, and dynamically computes a timeout for work units: exceeding this timeout will see the batch of links re-assigned to another worker. This system allows for great flexibility in workers, which may vary in batch size and connection throughput.

The storage manager is also responsible for enforcing atomicity of jobs and of the underlying corpus. This is enforced using a check-out system: workers check out batches of links that remain allocated until they are returned, completed, by the same worker. When a sample is complete, it

[16]The portion of LWAC used for parallel request management is available separately at https://rubygems.org/gems/blat

is 'sealed' on disk to prevent further edits, and the scheduler instructs workers to sleep until the next sample time.

```
1 root/
2 root/database.db
3 root/state
4 root/files/sample_id/sample
5 root/files/sample_id/1/2/3/456
```

Listing 3.1: Corpus directory structure on-disk

Data storage in the system is split between metadata, stored in an SQL database, and website sample data itself, which is stored as raw HTTP response data in a versioned structure on disk. The storage format is optimised for large samples, and is nested in order to avoid common filesystem limits (as shown in Listing 3.1). LWAC does not enumerate URLs in memory, meaning there is no hard limit on corpus size—instead it incrementally loads links as requested by the workers. This means that memory usage is limited to the maximum batch size within the system.

Workers are able to imitate the behaviour of end users' browsers as much as possible, so as to avoid search engine optimisation and user-agent detection tactics. They are able to provide cookies and request headers commensurate with a web browser, and follow redirects to similar depths. This behaviour is passed as part of the work unit itself, and so is configured centrally.

Workers are able to enforce limits on MIME type and file size, ensuring that documents are not downloaded if they are incompatible with further evaluation stages.

After downloads have occurred, data may be retrieved for analysis in a variety of formats using the included export tool. Export is possible at one of three levels: all data, individual samples (all datapoints at a given time), or individual datapoints (across all times).

Metadata is stored about network-level properties such as timeouts and latencies, as well as URL properties (link length, etc.) and HTTP headers. The original request parameters and sample design are also presented in the same structure, allowing corpora to be somewhat self-documenting. Appendix A shows the object hierarchy of page data.

Exports are performed by filtering data using a series of expressions. These data are then passed to formatting expressions, which are able to normalise display formats for analysis, before calling a number of pluggable formatting modules. Current formats supported are:

- CSV format with one-row-per-datapoint or per-sample;
- A collection of CSV files, split per-sample;
- JSON collections, for import into tools such as MongoDB;
- Arbitrary text output using the ERB template format. Supports server, sample and datapoint level exporting;
- XML, transformed from the original input parameters using XSLT.

Resource Usage

LWAC is designed to handle corpora of unlimited size, and uses a pipelined design to avoid enumerating its corpus members. In practice the limit on corpus size is imposed by the SQL server used.

Memory usage is $O(1)$ for both the storage server (which backs all of its corpus data with the disk) and the client (which only ever loads as many datapoints as number of parallel connections). Ultimate limits are defined by the batch size chosen for workers:

- Server: $(clients \times batch\_size \times link\_size) + (batch\_size \times max\_resource\_size)$
- Client: $in\_progress \times max\_resource\_size$ (using disk cache)

Workers use static quantities of disk, equal to the maximum download size of each page multiplied by the batch size. Disk usage for the storage controller is $O(N)$, though the storage format is well suited to deduplication using symlinks to point to previous data.

3.2.2 Performance

Ultimate performance is dependent on a number of factors:

- The number of worker machines used;
- The number of connections per worker;
- Worker network speeds;
- Latency of DNS and web servers.

The last of these implies gradual degradation of performance as links rot. Performance figures stated here are therefore for the 'best case' scenario where links are all available from fast servers.

Artificial

Figure 3.5 shows the download rate for 140KiB HTML files served over 100Mbit ethernet using the Nginx web server. This server configuration was designed to offer high-performance yet still be representative of common web deployments[17].

The performance of parallel downloads is largely constrained by the platform, in this case the Linux 2.6 kernel. Downloads are reliable until roughly 700 parallel connections are used, beyond which requests are dropped locally as half-open sockets are closed. Until this point performance is roughly log-linear, with 100,000 requests taking just 100 seconds with 500 parallel connections. For subsequent tests, 400 connections per worker machine were used.

Selecting the number of worker machines required for a deployment is a tradeoff between many factors, particularly cost and download rate[18]. As the server is responsible for all co-ordination, the speed improvements seen when adding worker machines is nonlinear, eventually

[17] Nginx is known for its ability to serve many users at once, and "[a]ccording to Netcraft, nginx served or proxied 21.64% busiest sites in May 2015."—http://nginx.org/en/

[18] Reliability and physical location are also likely to influence results, especially where geography affects the contents of pages

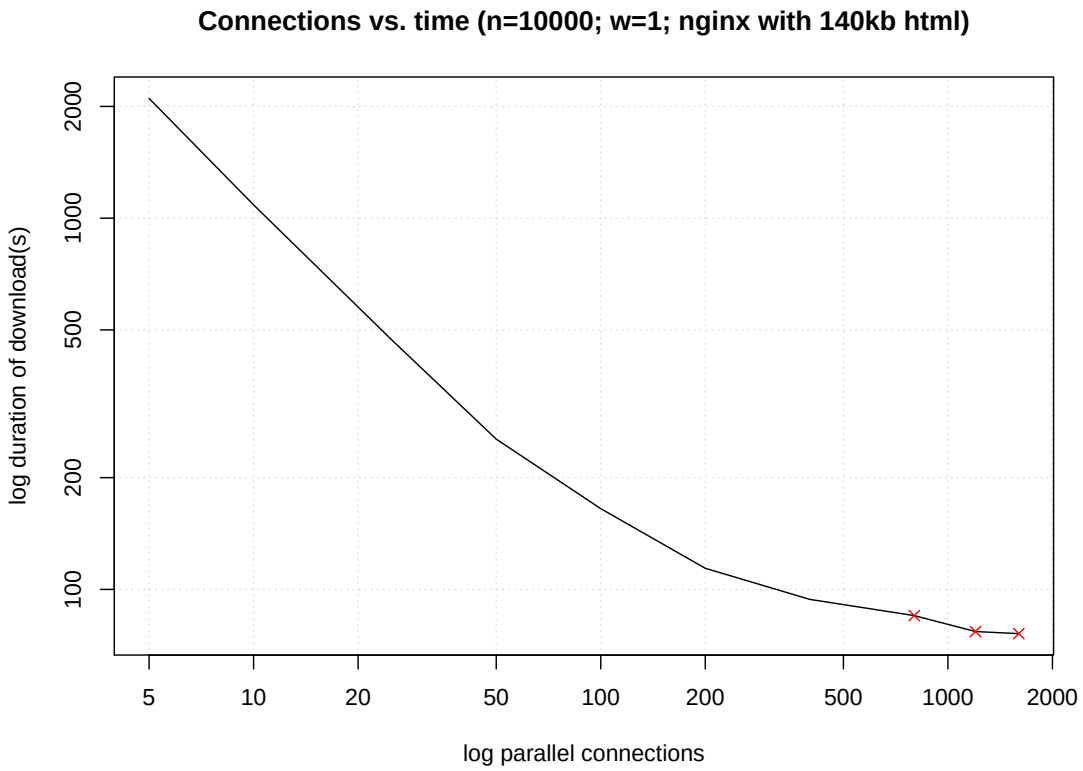


Figure 3.5: Download speeds for one worker for various levels of parallelism

plateauing where the overhead of the batch request equals the difference in time between requests made by n and $n + 1$ workers.

Figure 3.6 shows the overall download times for the same dataset when using differing numbers of workers. Workers were all connected to the storage manager via 100Mbit ethernet (minimising contention), and were making requests via the same link to a local test machine. As expected, the improvements soon plateau, though there is a large amount of variance—larger batch sizes and a larger test corpus are expected to show further improvement with $n > 6$. This plateau may be expected to occur earlier if workers are connected via relatively low-bandwidth links (for example, if they are connected over long distances via the internet or a VPN).

Real-world

In order to test retrieval performance in a real-world scenario, a test deployment was constructed consisting of three clients, each making 400 parallel connections. Each was running Linux 2.6 connected to the storage manager using a 100Mbit ethernet link, and to the internet using Lancaster University's tier-1 fibre connection. This setup was designed to be within reach of many academic users, who are likely to have bandwidth allocations exceeding the sites that they visit, yet have limited access to hardware.

The 228,000 URLs used were sourced using BootCaT, making 4600 requests to the Bing search engine and retrieving 50 links per request. This resulted in a corpus size of 14.9GiB, which, after

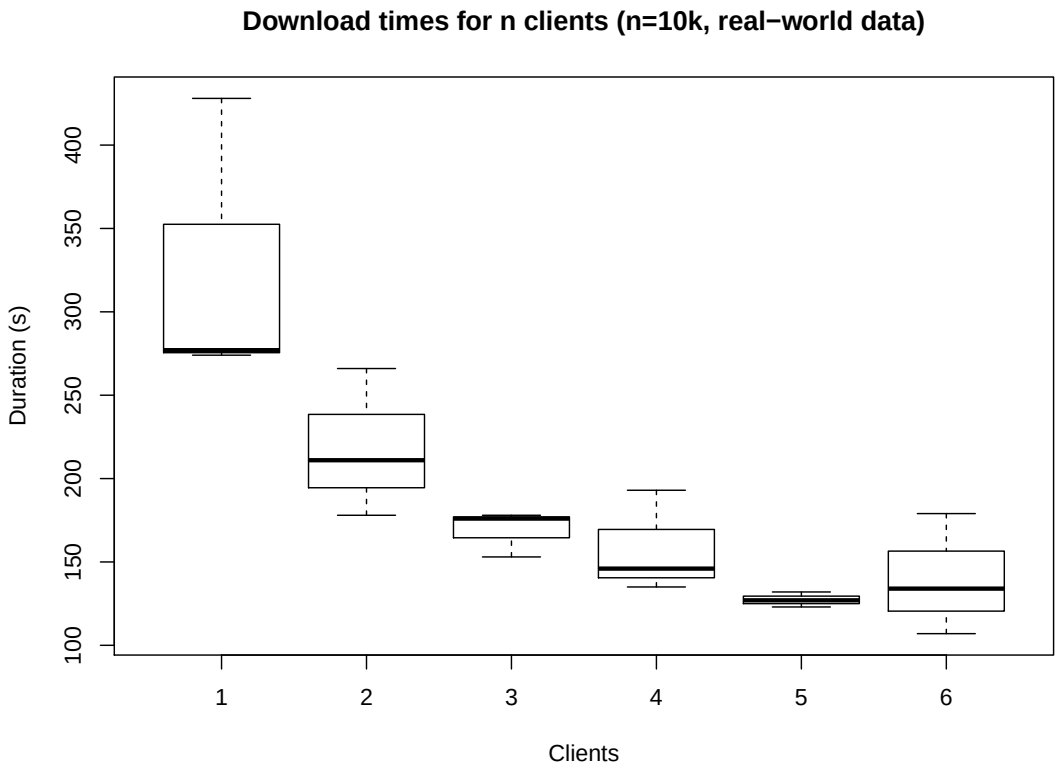


Figure 3.6: Download speeds for 10,000 real-world HTML documents.

boilerplate removal, yielded 588 million words.

The throughput over time is shown in Figure 3.7. This data was summarised by examining the logs of the storage manager, and thus represents rate calculated after each worker has checked in its batch—hence the pattern of jagged increases followed by gradual decay. Larger spikes are seen towards the left hand side due to synchronisation of the check-ins by clients: eventually variations in batch retrieval times cause these to disperse and smooth out. This deployment can be seen to plateau at roughly 80 links per second (roughly 288,000 per hour), meaning the full 228,000 page corpus was retrieved in roughly 45 minutes.

Using this deployment would mean that retrieval of a corpus the size of the BNC is possible in around eight minutes; a ukWaC[49] copy would take 2.5 hours (or 17 if using the pre-filtered url list), and DECOW2012[155] would take 24 hours.

Etiquette

Ordinarily, crawlers are expected to adhere to well-acknowledged guidelines regarding page retrieval, revisit times, and link traversal. LWAC ignores these by design, due to a number of factors that differentiate it from standard crawlers.

Firstly, the integrity of the returned samples is impossible to validate if they are not reliably retrieved. Any missing data would have to be imputed, reducing its scientific value and complicating the process of analysing results.

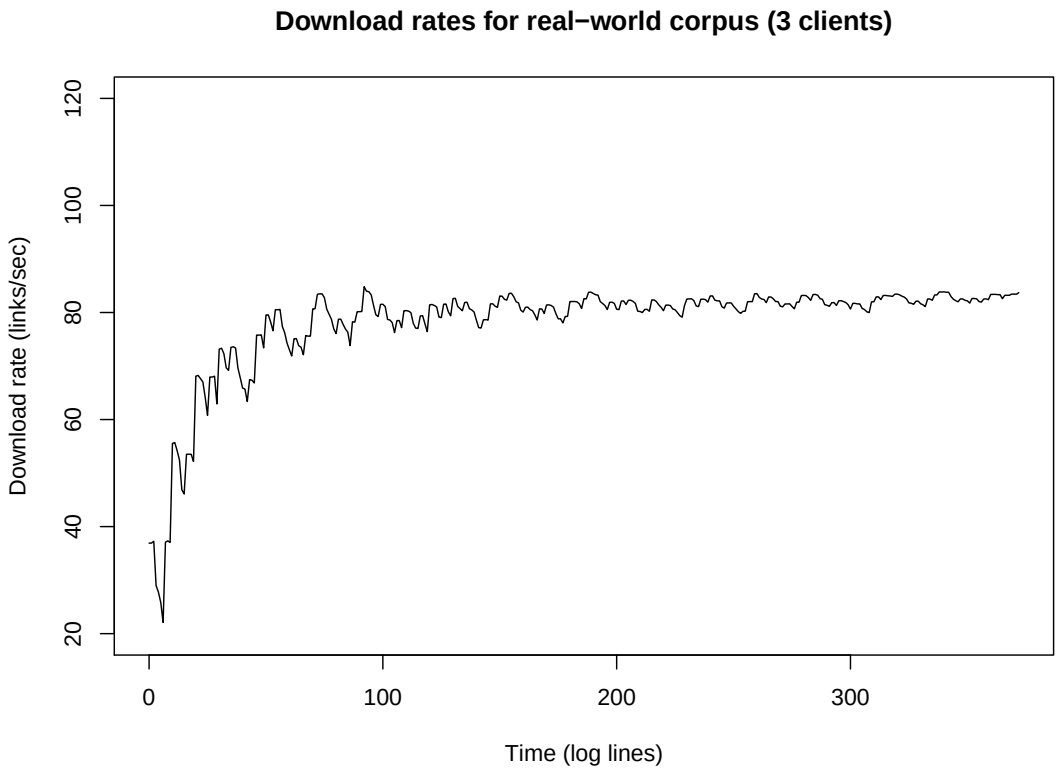


Figure 3.7: Download rate over time for real-world BootCaT derived dataset.

Secondly, though this is largely under the control of the user, it is expected that links will not be visited so regularly as to significantly increase load on a given server. Most effects requiring longitudinal analysis take place over a relatively long period of time that does not warrant short sample periods.

Finally, LWAC's link selection is unlikely to include many pages hosted on the same server. This is not guaranteed, however, for web-wide longitudinal analyses each 'hit' by LWAC is not going to result in immediate successive hits to the same server, as it would with a crawler that navigates pages. Where this is violated, it is of course up to the user to mitigate any load to the web server. LWAC's randomisation of the order of link traversal also helps mitigate this, as it reduces the chances of a set of links co-occuring systematically in each sample.

Though it is possible to perform DDOS-style attacks using LWAC, the use-case for which it was designed is unlikely to violate the principles adhered to by existing crawlers.

3.3 Summary

Temporal effects, particularly the loss of content over time, have been well documented for web data. The effects of this are relatively well understood relative to common technical variables, and this is widely used to improve crawling for search engine creation.

Though link rot (and other page content changes) are evident in web corpora, the effect this has upon the contents of web corpora is far less known, as few studies have focused on the

content of documents.

LWAC is a web corpus construction tool that takes a novel approach to retrieving data across time, using a cohort sampling design. Many more regressor variables are retained than existing systems, allowing for more in-depth regression, and a cohort sampling design makes for easy inspection using existing survival analysis techniques.

The performance of the tool is sufficient for longitudinal sampling of large, general-purpose corpora in the gigaword range. This ability can be scaled during runtime by addition of worker nodes, if necessary.

Data is presented by LWAC in formats that are designed primarily for quantitative analysis—this is well suited to the large volumes of data produced, which for many tasks will swamp human evaluators. A large number of regressor variables are also stored, making the resultant corpus suitable for studying not only questions regarding the content of pages, but also network issues such as latency and reliability of web resources.

A large-scale study using this tool (something that is beyond the scope of this thesis) allows for the quantification of effects that, in conventional corpora, have been informally managed. This may be used either to control for such effects when building corpora using web data (for example, to make document age more closely match offline or auxiliary data), or to analyse changes in classification over time: 'age' simply becomes another external variable, to be conditioned upon at will. This approach is revisited in Chapters 5 and 6, which provide methods for such parameterisation.

Chapter 4

Proportionality: A Case Study

The degree to which a corpus represents 'real' language use is arguably the most important aspect of corpus design. As discussed in Chapter 2, general purpose corpora are typically constructed based on a mix of expert opinion and per-medium data sources such as bestseller lists.

Reasons for this approach are both theoretical and practical—sampling individuals in the act of using language is very difficult, and the persistent nature of texts means that they remain available for sampling after such 'usage events'. This is especially relevant where the corpus designers intend to include older works.

Nonetheless, there is a missing empirical link between the expert-guided designs of language proportions within a corpus and the ground truth of an individual's (or population's) language use.

This chapter describes a case study where a census of language use was constructed for a single individual, described by genre and source. This sample design yields very little inferential power about whole population, but serves to illustrate the disparity between the proportions a general purpose corpus may have, and the language used by an example individual. This more closely follows Hoey's [75, p.14] model of describing the body of language a single user is exposed to:

> *Not even the editor of the Guardian reads all the Guardian, I suppose, and certainly only God (and corpus linguists) could eavesdrop on all the many different conversations included in the British National Corpus. On the other hand, the personal 'corpus' that provides a language user with their lexical primings is by definition irretrievable, unstudiable and unique.*

The extent to which this one subject may be a useful guide for corpus building is not addressed directly, but this is an area that could be expanded using demographic auxiliary data in order to fulfil at least some requirements of Hoey's personal corpus.

Though beyond the scope of this thesis, this approach could be generalised for use with questionnaires and other less-invasive data gathering techniques, or specialised and restricted to a given source of data for use with automated tools (for example, analysing only a user's web history). In addition to further illustrating existing sampling problems in corpus linguistics, the methods described herein offer some value for further work on lexical priming.

In Section 4.1, I describe the design of the sample being built in the study, focusing on the key differences between the approach taken here and contrasting it to Brown-influenced corpus methods.

Section 4.2 examines how the availability and ubiquity of technology has enabled people to mitigate practical issues with data gathering, and surveys the efforts of life loggers, whose goals often align well with those of this study.

Aims and objectives for this case study are covered in Section 4.3, and the sampling policy is specifically stated.

Section 4.4 describes how this sampling policy was implemented, and the design decisions affecting different technologies used to record data. Each data source is identified and categorised according to its persistence and ease of sampling. Later, in 4.4.4, this section focuses on operationalisation and post-processing the recorded data.

Results are shown quantitatively in Section 4.5, and the dataset is described with some comparison to the BNC. The data are broken down by linguistic events and by words, and summarised.

Section 4.6 analyses this data qualitatively, drawing comparisons against the BNC with relation to the subject's demographics and to the activities performed within the sampling period. This section also addresses how the data inform the method, evaluating key aspects of the data gathering process such as validity and ethical concerns.

Finally, we conclude the chapter in section 4.8 with a short review of the methods, and a discussion of how these contribute to the overall themes of the thesis.

4.1 Sample Design

As we have seen in previous chapters, conventional corpus building efforts centre around linguistic variables, and rely largely on expert opinion to balance their socioeconomic variables. This approach was initially selected in order to avoid certain practical problems (many of them caused by technological limitations), though it has also caused others, most notably the difficulty in retrieving metadata about texts post-hoc.

Often, the only way to ensure demographic balance is to rely on sources of auxiliary data such as bestseller lists and library loan records that index textual variables by socioeconomic ones. This process of using 'proxy' variables is particularly opaque, and often limits the metadata available to that in the list used for selection.

The design used for this chapter's census is instead based on observing 'linguistic events'—informally-demarked single uses of language—in an ad-hoc manner. This allows the context to be recorded and metadata to be captured without retrospection. In order to accomplish this, detailed logs were kept on everyday activities.

This approach is far closer to that used in some special purpose corpora, especially where the restricted domain allows use of automated recording tools or sampling from an existing rich database (such as from an online forum). Sampling in this way offers guaranteed external validity

for linguistic proportions, but may be considered just a single data point from the population of language users and cannot be formally generalised to other people. The purpose of this case study is in part to explore the methods that may be used in creating such samples.

Of particular interest are variables that are particularly challenging to sample:

- The age of a text when 'used';
- Other temporal information, such as the times and days when texts are used;
- The social context of text use;
- Attention paid to a text, and which portions were read;
- Proportions of text types used, especially representation of types that may be missing from other corpora such as greetings, billboards, product labels.

The ad-hoc sampling approach, applied to *all* texts, also allows inspection of the proportions of language types used—something that is estimated for general purpose corpora. Though the scope of this case study is necessarily limited, a wider 'language use census' for many people would be a robust empirical method for validating claims of representativeness in general purpose corpora.

A further advantage of this method is that the population may be rigorously defined, as data on the participants is available to whatever standard deemed necessary. This contrasts with the use of existing lists or repositories, which have been constructed with differing purposes and levels of documentation.

4.1.1 Difficulties and Disadvantages

Sampling language use ad-hoc involves changes in method that are practically challenging. To build a true general-purpose corpus, one would have to take data from enough people to cover the population required, and each would have to undertake a fairly intrusive procedure to do so.

Difficulties encountered when trying to sample large populations are well documented in the social sciences, and many techniques exist to mitigate common biases (such as weighted sampling[91]), however, these are largely beyond the scope of this chapter. It is notable that one solution in sociology has been to share data in a style similar to corpus linguistics, to minimise the costs involved in executing a high quality survey[5].

The increased 'focus on the person' of a primarily social design also raises a number of ethical issues, as it is increasingly possible to derive information not just about a general group's language preferences, but about an individual's (or a comparison between groups). This is the inverse of its value, but it is nonetheless worth considering as many study designs in linguistics and NLP need not work with such sensitive data. This issue is addressed after the discussion of methods and findings in Section 4.6.6.

Further, sampling text as it is used raises methodological challenges—how can we be sure that the language observed is still naturally occurring and valid? All sampling is going to compromise

on this, and the extent to which one values detail over interruption will vary by study design. It is the aim of this chapter to identify major challenges in this area.

4.2 Technology

Conventional corpus designs were chosen to avoid practical challenges that were existant at a time when computerisation was in its infancy. The application of new technologies to the problems of sampling offers a way to mitigate a number of these issues, making alternate designs possible.

Two main themes are notable in easing access to text in a usable form:

- Many more documents are now produced and consumed in digital form. This means they are accessible for automatic copying and processing by sampling software, often without any intervention necessary by the user. As techniques for cataloguing, monitoring, and annotating documents improve these data become richer, often in ways that would benefit an end user (and thus a corpus builder).
- The abundance and ubiquity of portable technology such as smartphones lowers the difficulty of many existing sampling methods such as audio recording or photography. Connectivity of these devices allows for easier movement of data and offloading of processing, even when in physically remote areas.

The former of these is well represented by the Web-as-Corpus movement in corpus linguistics, and many corpora include sections which have been sampled by distributing portable technology (originally tape recorders). Of particular interest, however, is the value that we may extract by exploiting both to form a coherent sequence of events, and using that narrative to inform corpus annotation.

One community that has been using such techniques extensively is that associated with life-logging. Life-loggers record, and often catalogue, their own activities and use of many different kinds of resource in everyday life for reasons of posterity, entertainment, or memoisation.

4.2.1 Life Logging

A number of technologies have been developed as a result of the life-logging community's interest in multimedia records. Life-logging is an activity that emerged slowly out of the principles of webcam shows and reality TV, and involves recording (and usually broadcasting) continuous information about one's life as it occurs.

Though most popular efforts started as a means for providing entertainment, methods used soon diversified and gained the interest of the information retrieval and processing communities. Many projects have been started with an aim to catalogue and operationalise the huge stream of data each person creates, largely with a focus on aiding that person in their daily life, or aiding large organisations (such as defence forces) in management of resources and people.

Life-logging as a distinct activity is often considered to have started with Jennifer Ringley, who started broadcasting her entire life using webcams in 1996 ('JenniCam'[19]). Her website proved successful for many years, gained significant media coverage, and in many ways can be credited with popularising lifecasting (broadcasting of one's life rather than simply logging it) and helping to define the most recent wave of life-logging efforts. However, she did not start the practice of either life-logging in general, or life-casting.

Life-logging using less technical methods has arguably been performed by millions in the form of diaries. Though informal, the value these can offer as databases is well-known to historians as they offer a narrative structure that is difficult to build from other sources. The case study detailed within this chapter uses such methods for exactly that reason.

More formal, detailed forms of diary could properly be called the first life-logs. Of particular note in this area is the 'Dymaxion Chronofile'[53]—Buckminster-Fuller's attempt to document his own life, which consists of a series of scrapbooks recording his actions (and documents) every 15 minutes between the years of 1920 and 1983.

Buckminster-Fuller's efforts captured the last 63 years of his life in extreme, multi-modal, detail: capturing all correspondence, bills, personal notes and material such as clippings from newspapers. This detailed record of his life is essentially unmatched in the pre-digital era, and would not be attempted again until the digitisation of many common tasks eased the process of capturing documents.

For the first lifecaster, we must turn to Steve Mann. Mann began working on wearable computing in the early 1980s[123], focusing on video recording and head-up-displays, however, it is clear from photographs of his equipment that it could hardly be considered unobtrusive enough for sampling purposes (indeed, it would probably prove more troublesome than Buckminster-Fuller's 15-minute interruptions).

The posterity-oriented and academic streams of life-logging persue a parallel course from the mid nineties onwards. The aforementioned success of JenniCam led to a number of copycats, and, eventually, services such as ustream[82] and Justin.tv[81]; both of which make lifecasting available to anyone owning a smartphone. The consumer side of life-logging has also been rising in popularity due to products such as Narrative[108][20] (which emluates Microsoft's SenseCam[74]) and Google's Glass project (which, though not explicitly designed for life-logging, provides hardware well suited to it). There is also an increasing interest in smart watches, which promise to further extend the simplicity of accessing smartphone features in a continuous, unobtrusive manner.

Academically, much more research was undertaken on uses of the data gathered. This meant a greater focus not just on multimedia methods and streaming, but also other sources of data such as documents (digital or paper), location data, etc. Microsoft's Gordon Bell is particularly well known in this area for developing SenseCam[74], which focused on photography and location data, and MyLifeBits[55, 54], which consolidated many different sources. In these cases (and others' similar efforts[79, 41, 39]), the focus was on producing a record of life that could be used

[19] www.jennicam.org, currently down.
[20] Originally known as 'Memoto'.

by the original subject as a form of super-accurate memory, similar to Vannevar Bush's 'Memex' vision[32].

In one study, mechanisms for de-duplication are examined with a view to improving the interface used when browsing large collections of logged data[56]. Therein Wang & Gemmell develop a filter based on clustering that identifies not just identical documents, but also those similar enough to carry identical information in slightly different forms (for example, with updated boilerplate features or layouts):

> *. . . we count both the exact duplicates and near duplicates we find, we can effectively reduce 21% of all documents and 43% of web pages from the viewer to reduce clutter of the user interface.*

A similar project, funded by DARPA (but later dropped) was simply called LifeLog[3]. This took a similar approach to MyLifeBits, aggregating many sources of data into a single bundle, organised by time. Unlike MyLifeBits, however, much of the focus was on using information retrieval methods to construct a coherent narrative for a person, which could later be interrogated at a higher level than the collected data itself.

Work on 'audio-based personal archives' by Ellis & Lee[43] moves the focus of much of this logging from visual to audio recording, with the intent being to create a digital memory of events that is based on the most inconspicuous and unobtrusive recording methods. They present a number of audio analysis techniques, including those for anonymising data without loss of other information, that may be used for speech data.

Much of the focus of these projects was on recall and operationalisation—tasks that are particularly difficult and largely ignored by the more entertainment-focused communities. It is unfortunate, however, that their purpose was largely one of narrative creation: many exploration and recall tools rely on a human's ability to interpret the complex data recorded, and this is ill-suited to the more formal sampling approach needed here.

One notable exception to this is Deb Roy's project to record the language use of his own child (most famously exposited in his TED talk[151]). This involved continuous video recording using a number of cameras in his house. The footage from this was then analysed to identify regions where his child was speaking, and thus develop a corpus for use in language acquisition studies. He used a number of methods to identify interesting video regions that, it was initially hoped, would apply to analysis of the speech portion of this case study's corpus. In reality, however, the conditions under which he records audio proved significantly simpler than those in this case study.

Because of the continuous nature of life-logging, efforts have been made to use methods that are easy to maintain, self-contained, and covert. Due to this, as well as the original intent of the life-logging process, much of the effort surrounding life-logging focuses on multi-media sources, and how they may be best combined to form a coherent idea of context.

Typical sources of data considered include:

- **Video recording**

Many life-loggers wear systems that are able to continuously record video in the direction they look, and upload this using mobile networking systems.

- **Audio recording**
  Due to its lower obtrusiveness, many efforts surround the analysis of audio logs, and include systems to detect voices and identify events such as making appointments.

- **Document storage**
  With the increase in use of born-digital documents, some of the more holistic life-logging systems record documents as they are read, with a focus being to integrate this data into the larger picture. Others scan in physical documents such as their post for later retrieval.

Many of the requirements of a life-logging platform (covert operation, comprehensive data management, context identification) overlap notably with the methods used in covert sociological research and, of particular note for our purposes, those constructing spoken language corpora. The distinction drawn here is one of philosophy. Where life-logging focuses on construction of a narrative, contextualising data, more conventional sampling methods have focused on data recording and normalisation in a formal setting, discarding data that is not immediately operationalisable.

Notably, one of the methods used to create the spoken portion of the BNC was covert recording, where 124 people were provided with tape recorders[30]:

> *Recruits who agreed to take part in the project were asked to record all of their conversations over a two to seven day period. The number of days varied depending on how many conversations each recruit was involved in and was prepared to record. Results indicated that most people recorded nearly all of their conversations, and that the limiting factor was usually the number of conversations a person had per day. The placement day was varied, and recruits were asked to record on the day after placement and on any other day or days of the week. In this way a broad spread of days of the week including weekdays and weekends was achieved. A conversation log allowed recruits to enter details of every conversation recorded, and included date, time and setting, and brief details of other participants.*

As illustrated by the BNC's demographic balancing of that portion of their corpus, this ability to directly record data from the field satisfies the disadvantages of text-index-oriented methods of document selection, allowing us access to all of the contextual data at the time of text consumption/production (this is particularly advantageous for spoken texts, where production and consumption often occur soon after one another).

The cost of this approach is (as felt by all sociological studies) a need to find a sufficiently large and heterogeneous sample of people who may wish to record data about their language use (and for long enough for it to be useful to researchers). This is arguably more difficult in practice than text-index-based methods, and should only be considered at a large scale where the difference is likely to be crucial to a study, nonetheless, large samples (such as the British Household Panel Survey[173] and the British Crime Survey[76]) exist in sociology as testament

to the value such designs may yield and the illustration that no alternative method exists for many sociological issues.

The ability to specify the demographic variables of a sample directly makes techniques using logging particularly applicable to the construction of special-purpose corpora, especially where those corpora are best demarcated along social lines. Indeed, at one extreme of this scale exists the concept of a personal corpus; something that may yield insights or models about a single person's current language usage. Such a resource may one day be particularly valuable in defining how one uses text-based interactive systems (such as the web) or reads content (such as news articles).

Such designs exhibit a tradeoff: a decrease in socioeconomic breadth in exchange for an increase in linguistic breadth. It is the intent of this chapter to illustrate the value in this approach, and to investigate methods by which it is possible to construct corpora without undue difficulty through the use of life-logging methods.

In practice the distinction between life-logging methods and 'conventional' data recording for research is one of rigor and intent. Life-logging is primarily concerned with capturing *all* data, often in a best-effort manner, using whatever means is necessary to unobtrusively do so—favouring imperfect but complete records. Conventional data capture techniques, though often overlapping methodologically, focus on specific research questions; seldom recording data that cannot easily be operationalised and valuing quality recording of few variables, over opportunistic recording of many.

4.3 Aims and Objectives

The case study described here is an attempt to assess the extent to which techniques from life-logging may assist corpus builders in creating a demographically-oriented corpus, and to identify the extent to which the subject's distribution of language use supports the assumptions made by designs for reference corpora. It follows an iterative design in order to gradually refine the methods used, focusing on:

- The variables that may practically be recorded about a text (and which must be before they are lost);
- Methods that may be used to sample text unobtrusively, especially how new technologies may be used to assist;
- Methods and tools for operationalising logs after sampling is complete, and how these may ease the process of data gathering itself;
- How to minimise the intrusiveness of capture methods both to the experimentor (meaning that he can capture smaller interactions) and to those around him (meaning the data is more representative).

Ultimately, the differences in sample design contribute to a larger picture that could yield much future work. The issues addressed here are:

- What methods may be used for gathering data in a short-term language census?
- How may the collected data be operationalised?
- What proportions of language are used by the subject; do these support common claims from general purpose corpora?
- How may these methods be used in future to aid those building corpora?

The case study described here is a first effort in exploring a method that may be useful to many fields. Aspects of the life-logging approach to determining corpus properties could be generalised (or relaxed) in order to further reduce the demand on the subject. The scope of this thesis does not permit further study on these methods, however, it is written with a view to clarifying further work into questionnaire-based or electronic elicitation of language proportions.

4.3.1 Sampling Policy

The aim of a personal corpus is to emulate, as closely as possible, a census of observed language. Use of language, in this context, is defined as any conveyance of information, spoken or written, in any quantity. There are no bounds to the context in which it is used, nor the language itself, as the purpose of the study is to evaluate these very things.

Each of these transactions is recorded as a single line in the data set, and will be annotated with the variables recorded. One of the major issues encountered in preliminary tests was annotation for attention and proportions read. These will be recorded along with textual properties to ease operationalisation.

Attempts were also made to record sufficient data to retrieve the full text of each transaction. This worked better for some data sources than others, and much of the case study thus concerns itself with (sometimes estimated) word counts. Word counts were chosen as a measure of size due to their use in other corpora, and their applicability to many different media.

A number of practical challenges were identified before sampling, and these were backed up by experience:

- **Review and Production**
  Both should be recorded as fully as possible. Where a text is re-read, or developed and continually re-read, this should be noted as an ongoing process (and accordingly oversampled).
- **Short Utterances**
  Very short interactions, such as passing greetings, should not be under-recorded. Their inclusion is likely to be one major difference from conventional corpora.
- **Oft-reread Texts**
  such as labels, signs and the like. It's debatable whether or not one actually reads or merely remembers/recognises these.

4.4 Method

As this case study is exploratory, seeking to drive and refine the methods used for recording *all* language use, an iterative design was chosen. This saw a number of preliminary sampling periods, with a review after each to identify the strengths and weaknesses of each.

The subject was myself—this was done for a number of reasons: Firstly, legal and ethical issues surrounding recording and review of the data were mitigated by having the analysis performed by a member of the original conversations. Secondly, iterative review of methods involved was possible with internal, 'white box' examination of how data were collected, and what edge cases and procedural difficulties arose. Thirdly, the demographic status and other person-related variables are well known and need no formal elicitation, minimising time spent on construction of questionnaires etc.

4.4.1 Data Sources

Before the first iteration of the sampling/review process, all of the possible language data sources used in everyday life of the subject were informally identified. It became clear that these data sources exhibited properties that would make sampling easier, or less intrusive. They were classified by the methods required to capture their text:

- **Persistent**
  Resources that exist in a format that is immutable and easily retrieved. This covers many physical items such as books, and some broadcast media as well as notes made in a notebook. Only identifying information must be stored during sampling itself, in some cases merely an ISBN or similar index code.

- **Ephemeral**
  Language data that cannot be accessed after-the-fact in any way, or may differ by time or context. This most obviously contains speech, but also many websites, things such as billboards that cannot be readily accessed, or todo lists that get destroyed.

- **Digital Origin**
  Documents that are read, or written, on electronic devices. These may fall into either of the above categories, yet as they may usually be copied with no overhead, it is often simpler to store them at the time of use. Many document types are now digital, as well as the obvious sources such as email or online chat.

This classification was useful in order to minimise the intrusiveness of a collection method, whilst maximising the detail recorded for a given source (ideally to the point of storing verbatim text). In practice, methods were easy to develop for automated recording of digital documents, and many techniques exist for sampling non-digital persistent and ephemeral sources with scientific levels of accuracy already.

Sources initially identified by introspection are listed in black in Figure 4.1.

The inadequacies of introspective methods to comprehensively identify data sources soon became apparent during preliminary tests, as the process of recording brought increased

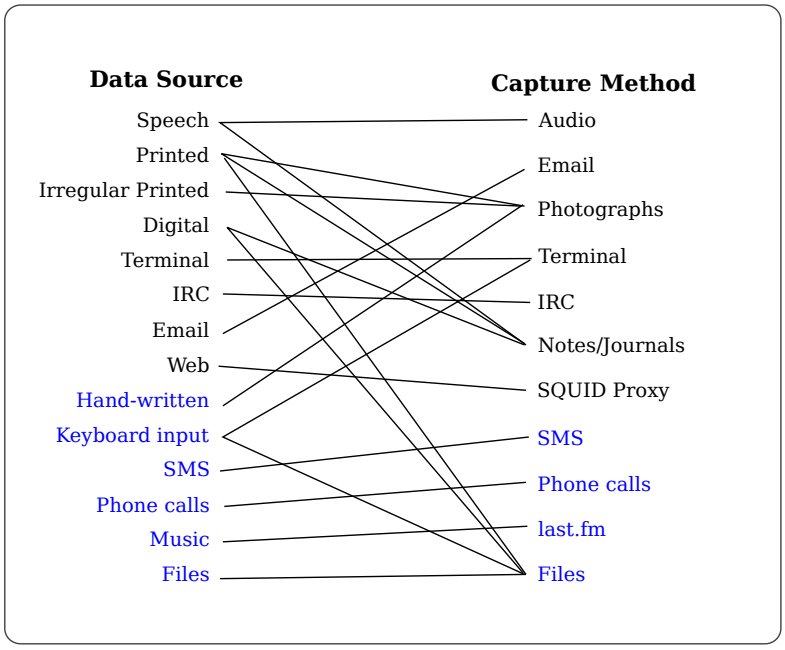


Figure 4.1: Data sources and their appropriate capture methods (those in blue were added during preliminaries)

awareness of language use, raising a series of edge cases. These were collected and used to augment the list of sources shown in Figure 4.1. As well as covering the set of sources to be gathered, methods of collection were chosen with flexibility in mind in order to cope with un-envisaged sources of data[21].

The list is necessarily furnished according to the life of the subject in question—from this study I am unable to assert that it is generalisable to others, though the process of doing so would involve relatively little intrusive sampling.

Each of these sources is 'covered' by one or more sampling tools. These tools progressed most during the iterative process, and each was subject to a number of procedural subtleties that were refined throughout the study.

4.4.2 Recording Methods

Methods used to record data were chosen for a variety of reasons. They must, in sum, cover the sources mentioned above, be unobtrusive both for the subject and those around him, and be sufficiently flexible to cover unforseen contexts and data sources.

These methods can be separated further into two groups: many methods are capable of recording multiple sources, and serve to form a narrative that describes the metadata of a linguistic event, pointing at another source for the data itself. These methods were chosen to allow for post-hoc sensemaking and narrative creation, something that was added to the experimental procedure after the first iteration indicated how difficult to operationalise much of the data would be.

The second set of methods are focused on a single data source, typically requiring little to no

[21] This flexibility has the unwelcome effect of slowing down analysis later on, and may be undesirable in some use cases

manual intervention to record data. Their records are either indexed by time, or by the more flexible methods mentioned above.

Indexing and Overview

Figure 4.2: The live notebook

Journals and Note-taking Two journals were maintained throughout the sample. The former of these was an A6 notebook maintained 'live' as events occurred (pictured in Figure 4.2). This was used to store durations of conversations, titles of persistent sources, etc.

The second was a nightly ('dead') journal, maintained at the end of each day in a narrative style. This blog-like record was intended to reflect in depth on the proportions of text used in each source, and how attentively each linguistic event was engaged in. The writing of the journal itself was not logged by any other methods. It was also possible to attach daily records to this journal, and the process of writing it inserted an opportunity to reflect on the mnemonic codes used during the day. This process is described in context in Section 4.4.3.

The live notebook proved to be the primary indexing method for all other sources of data, and its maintenance was the primary overhead of the study. As illustrated in Figure 4.3, each entry in the notebook was eventually reduced to a compressed form that roughly followed one-line-per-event, storing the time each event occurred, any identifying information deemed necessary for later memory of it, and a duration or other index of word counts.

Problems of simultaneous events and split attention were solved in the notes by having a start/stop event for ongoing events, and by using the nightly journal to reflect upon each event.

Audio Recording Following work on machine listening, the original intent of audio recording was to capture the occurrence and duration of conversations, as well as any smaller interactions

Figure 4.3: A page from the live notebook, detailing the format used

that would otherwise be difficult to capture (such as greetings, thanks when opening doors, etc.)

Capturing was performed with an Olympus VN713PC dictaphone, storing audio on a suitably sized external card that yielded many days' continuous recording. Provision was made to download recordings each night and store them with the nightly journal, however, in practice they remained on the recorder until the end of the study.

Aligning the recorder's output to the events mentioned in the notebook was a tricky process—Though the recorder itself supports index marks, there is a limit of 99, which was deemed too low for continuous use. Two devices were built to insert absolute silence onto the recording (something that is rare in real life and thus easy to programmatically detect), the latter of which is pictured in Figure 4.5. These clickers were to be pressed at the beginning of each conversation, so that voice activity detection could be performed to estimate the word count of each conversation (or, in an ideal world, extract verbatim text).

In practice, the process of tapping the button proved intrusive and, from the perspective of one talking to the subject, suspicious.

The mechanism used for the final iteration of the study was far simpler—the recorder's start time was written in the live notebook, and entries therein were keyed by computing the offset between the two times. Though this incurs a minor overhead in coding the data, it also allows for spontaneous conversations without much overhead, something that is particularly important to the study of text type proportions.

Photographs The primary method of capturing ephemeral, irregularly formatted, non-digital texts was photography using a cameraphone. This method was chosen largely because the ubiquity of smartphones in British society has led to a situation where photographing fairly

Figure 4.4: The audio recorder used

mundane items is widely unquestioned.

The smartphone used, a Motorola Milestone, also stores time and location data in its photographs using EXIF tags (as well as storing photographs sorted by day). This metadata meant that there was often no need to file an entry in the live notebook, and the cameraphone could simply be used in a very unobtrusive ad-hoc manner.

In earlier iterations of the study, it became apparent that the loud shutter noise made by the Android operating system when taking photographs was problematic. Though photography of signs, packets and such remained unchallenged, the attention of people nearby was drawn to the 'weirdo with the cameraphone' all too readily. This was solved partially (and with great difficulty) by disabling the noise, though it was still apparent from posture when a photograph was being taken.

There are notably a number of products available that continually take photographs for the purposes of life-logging. These were considered for the study, but their aims are generally to capture each event, rather than specific aspects of selected scenarios. The ability to consciously specify that the subject was more attentive in some situations (and take pictures accordingly) was judged to outweigh the value of having a continuous record (something much more capably performed by the journals).

Targeted Methods

Phone Calls & SMS Messages Both of these are automatically logged by the Android operating system used by the subject, and each was also indexed in the live notebook. The data was extracted using a free application that exported to XML.

Figure 4.5: The second audio index marker

IRC IRC was logged by construction of a bot. This bot accompanies the subject into chat-rooms and logs all messages observed, applying a rough human interest model to ignore data encountered when the user is set to away.

Web The SQUID webproxy was configured to log all traffic, and a number of logins were provided—one for each of the subject's internet-enabled devices.

The logs from SQUID store all requests, including advertising/tracking calls, downloading of things never read by the user (i.e. CSS and Javascript) and AJAX calls to partially reload pages. As such, a large amount of processing was necessary to extract URLs from these logs, and to parse the resultant data into a usable format.

Keylogging/Terminal recording Terminals and keyboard input were logged using a custom application that wrapped a terminal, recording the time each character was sent or received to/from the shell.

Each terminal created started recording to a new log file, storing the time at which it was started and a series of offsets from this time.

Last.fm In earlier iterations it was apparent that lyrics in music were being missed as a source of text—all devices capable of playing music were configured to 'scrobble' to the last.fm music service during the sampling period.

Though last.fm do not make their data freely available for access, third party tools exist to scrape their website and download detailed logs of tracks listened to.

Files Files were identified in a number of ways.

Some, particularly those on which the subject worked and contributed data, were written down in the journals. This is a precise method of separating what has been read, but incurs a large overhead.

Since the subject works entirely on projects and files that reside in a repository managed by a Revision Control System (RCS), the logs from each commit were used to generate a diff, and this was accessed after the sampling period to identify contributions made.

Another policy that may be used is identification of files by unix mtime (modification) or ctime (inode change, often creation), however, this is fraught with inaccuracy, as files are liable to be modified on disk many times whilst being edited, and sampling the differences is likely to happen at haphazard times. Further, this technique would capture many log files and others that have been edited by processes where the subject was not involved linguistically. By contrast, commits to an RCS are scheduled around logical additions, and are manually pushed so that only deliberately edited files are stored.

Files uploaded whilst on other systems may be uploaded directly to the nightly journal, or stored on a flash drive that was carried specifically for the purpose. In practice this did not occur during the recording period, though experience suggests these contingencies would be necessary if a longer sample were taken.

Email Emails are, again, stored automatically with sufficient metadata as to make them self-documenting. However, rather than presume all were read in a given day, each was tagged after being read with a label corresponding to the day.

At the end of the sampling period, these tags were collected and downloaded in mbox format, whence they were processed by the operationalisation script.

4.4.3 Recording Procedure

The study is based around a period of continuous sampling using the methods discussed above. For two weeks (in some cases longer), data were captured for each source. This process was structured around a daily routine:

Upon waking, and before any language was used, the recorder was turned on and a note of the time at which this occurred was made in the live notebook. Recording was then continued until the end of the day without interruption.

The live notebook, smartphone, and flash drive, were carried at all times. Since each of these could be backed up (the smartphone even did this backup automatically), the most data at risk was a single day.

Notebook entries were made as soon as was possible without interrupting the linguistic event being recorded.

At the end of the day (immediately prior to sleep), a journal entry was written in the nightly journal, and SQUID logs were uploaded for the day. This journal entry forms a narrative, estimating the time taken and attention paid to items in the live journal for that day, as well as detailing anything that may be written in shorthand-mnemonic form.

4.4.4 Operationalisation, Processing and Analysis

Normalising and operationalising such heterogeneous data without large overheads proved to be a significant problem that was only partially solved, and the data set presented here required manual editing that was possible in part due to the fact that the analyst is the subject.

This advantage, clearly, cannot be relied upon in other studies, and this part of the method demands most further study in order to define typical parameters for many processes that are dependent on human properties.

Two main processes were followed in data processing. The former of these was aggregation and normalisation—each data source was collected and transformed into a one-event-per-line CSV containing a standard set of fields (the selection of which was modelled on Lee's BNC index[110] in order to facilitate comparison).

After this normalisation process was complete, data were manually annotated to complete any fields that were not stored in the original metadata. This was largely an objective, uncreative task that simply demanded human reasoning capacity, but it is inevitable that some bias will creep in at this stage.

The second stage of processing involved coding text types and roles. This task is altogether more flexible and subject to design errors and bias than many of the normalisation stages, and was thus attempted in a manner that was designed beforehand. Since the aim of this study is, in part, to identify text types not seen in other corpora, following an existing taxonomy would necessarily limit the coding of any newly discovered. True free coding, however, is likely to draw distinctions between text types that are not made in existing taxonomies, rendering them incomparable.

The process followed was a hybrid approach—data were freely coded by inspection of the texts, but this was done with deliberate prior knowledge of Lee's classification scheme. The intent was to categorise texts according to Lee's scheme only in so far as they were deemed suitable by the analyst (who is also, lest we forget, the subject).

Though this approach was suitable for the aims of this particular study, it is difficult to advocate for any others using the sampling techniques described, and its use here should not be taken as such.

Human Interest Models

Beyond coding, by far the largest single influence on the data recorded was the human interest model applied. This was created in order to take into account two factors that had become particularly apparent (and notably do not apply in the same manner to conventional corpus designs, where many eyes may cover a whole document in sum):

- Often, only small (usually predictable) portions of a text are used. For example, I have *started* to read more books than I have *finished* reading. Generalised, this means that even Brown-styled corpora should favour the start of their texts slightly when selecting excerpts. Some media were more susceptible to this than others, and the automated normalisation

tools were built with facilities to take this into account[22].

- Texts, especially broadcast media and speech, were often used whilst also accomplishing a non-textual task (or sometimes both at once, such as talking with the radio on in the background).

Both of these were noted in journals, and added to the processing toolchain—each data source's normalisation script contained a model to extract the portions of text that were read, and each row of the normalised data format contained an 'attention index', ranging from 0–1, that served as a coefficient of the word count.

Though crude, this measure was able to produce approximations for word counts that were inline with the expectations of the subject. It is recognised that this may not hold much scientific value to others wishing to replicate the study, and in general it is necessary to investigate the inter-person variability of these properties in order to create more generalised processing tools.

Web logs The human interest model for web logging was built by inspection of the web logs and cross-referencing with information about the hosts identified. The normalisation script is concerned primarily with removing material that was downloaded without ever having been viewed, for example, non-text MIME types and advertising, and contains a multi-stage strategy for excluding content:

1. Filter only successful requests;
2. Filter only those requests that are of textual mime types (`text/*`);
3. Apply a blacklist of advertising websites and file extensions (manually constructed);
4. Discard links where the page was reloaded and the URL is the same as the previous entry in the list (this pattern is often caused by initially connecting to the proxy, for example);
5. Discard Non-visible text items (non-body elements if the file is HTML), and remove markup.

Beyond this initial normalisation, a spreadsheet's `LOOKUP` function was used to manually assign attention coefficients to the domains, based upon entries in the journals and interesting portions of web pages that follow regular structures.

Days were classified as changing at 4am, since there were no points in the data set where the subject was still using the web at this time.

IRC Logs IRC logs were already stored using a limited attention model, which was based on the principle that, in IRC, conversations are started, live for a short time, then die off as people in the channel return to work. The bot performing logging would start logging (and continue to log) for as long as the subject was talking, stopping 10 minutes after the final utterance.

In order to capture a human notion of day, that is, one demarked by sleep rather than midnight, the start time of each conversation was used to determine the day its data fell into.

[22] Amazon have the ability to monitor which pages are read on their Kindle ebooks, meaning that they possibly have a large dataset related to this effect.

Terminal (Console) logs Terminal data was logged with timestamps on each individual character, and the model was thus responsible for inferring when a command had output, and suggesting which portions of text were still on screen.

This was done with a 'timeout' and a 'scrollback'—the former describing a delay that had to be present for the text to have been read (rather than simply scrolling offscreen), and the latter describing the average size of a terminal (and thus the number of rows of text that remain displayed).

These parameters were tuned to match the specific data sampled over the period—for example, the recording period included terminal use displaying logs from software that was being developed, and these would produce thousands of lines of output before a pause.

4.4.5 Coding and Genres

Since the case study is in part aiming to identify genre distinctions not found in other general-purpose corpora it is not possible to select, a priori, an authoritative taxonomy of genres. This problem is further complicated by the mix of spoken and written data in the corpus, something that would usually be more formally separated.

Genre distinctions were made using a process that loosely follows the principles of the coding stages of grounded theory[59]. This involves a focus on free coding, and a consideration of all available context and information—something aided by the 'memoing' innate in the design of the journals. Free coding was performed with prior knowledge of Lee's genre categories.

This process was done in order to deliberately apply distinctions made by Lee where these seem appropriate, but to retain the flexibility to deviate where portions of the corpus did not fit comfortably within the existing categories. This soft alignment was chosen in order to ease comparison against existing corpora, especially the BNC. No effort was made to adhere to Lee's specification directly, so as to prevent forming a disconnect between the Lee-inspired categories and those that were sufficiently different to form their own.

Because of the large volume of data, and the manner in which it was extracted from its original sources, coding was performed using a semi-automated process on a source-by-source basis. This was chosen in part because the source itself refers to how the language was used by the subject, rather than simply what form it was available to the corpus builder, and is thus a contextual factor affecting genre distinctions. This would have to be documented in more detail for other studies.

For data that were processed and annotated largely automatically, codes were assigned after a systematic review of the data itself, using a spreadsheet to reduce the number of lines according to variables that were assumed to define a given genre (for example, web data was split by domain, and music was assigned by artist). These distinctions were then written into the extraction tools as heuristics, or applied directly before being merged with the main corpus.

For data such as images and notebook entries manual coding was required. This was completed using similar tools, except that it was seldom possible to reduce the data set before coding, and thus was not possible to apply heuristics to impute data. Data recorded in these

formats were often more varied with many unique or esoteric entries (such as the single tax disc[23]).

In order to better align the informal genre distinctions with Lee's format, the final document genre distinctions were formed by prefixing the data source, i.e. transforming 'rock' into 'music/rock'. These resultant genres are used herein for describing the corpus, as the basis for any direct comparison with Lee's BNC index.

This process of augmenting manual entries method would, if extended to use external data sources (such as the CDDB music information service[24]), be capable of automatically assigning genres for a number of data sources, greatly easing the manual intervention required. It is also likely that improvements in computer vision (such as Project Naptha[106]), or applications of existing databases (such as Wikipedia, or the web itself) could be used to extend these methods further to cover images and other data that were manually processed here.

4.5 Results

These results describe data collected during the third iteration of sampling, which lasted roughly two weeks during April 2013. Table 4.6 describes the availability of data per-day.

The period marked from Friday the 5th to Friday the 19th (inclusive) forms the sampling period described herein (with exceptions to this noted where made). This was the period for which the live notebook was maintained along with other intrusive data collection techniques, though notably some methods continued afterwards due to their ease-of-use.

Audio data is missing for the 19th as sampling was initially intended to end the day before, however, for producing estimated word counts this was not used as other recording methods cover the data sources. Though attempts were made to use the audio data, difficulties in automated processing meant that transcription (or precise word counts) were only possible using manual review (something only the subject could legally perform). Word counts were instead based on an estimated words-per-minute.

This fifteen-day recording period covers 8,619 linguistic events, encompassing an estimated 980,000 words. In total 3.4GiB of data were collected, though the majority of this figure (2.3GiB, 67%) of this is audio data. This is roughly in line with other life-logging studies[54].

The overall word count is changed massively by disabling the attention model, rising to roughly 5 million words. As we shall see later, this is largely because the data sources and genres which dominate the data set are particularly heavily adjusted. This raises the question of how we may ensure accuracy of attention models, something that shall be addressed in the discussion section below.

Variables recorded on each linguistic event are based heavily on Lee's BNC index, augmented to take into account the attention index and word count estimates. A summary of these is provided in Table 4.1. The `missing` column describes the proportion of rows missing this field,

[23] A format that now no longer exists after the UK's move to managing vehicle excise duty electronically.
[24] Now known as 'Gracenote': http://www.gracenote.com.

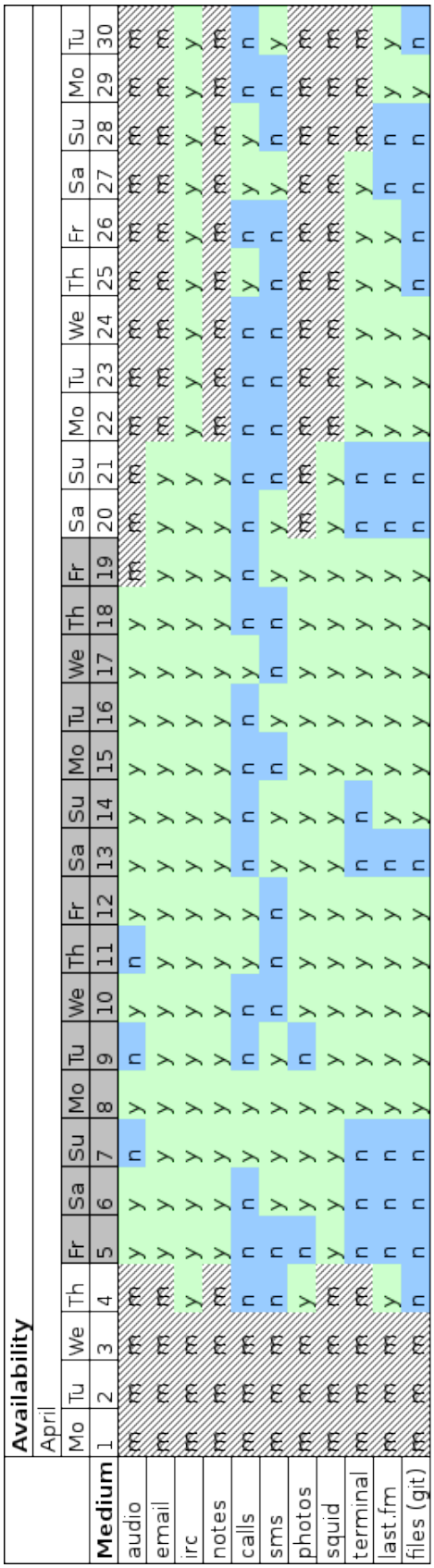


Figure 4.6: The availability of data from various sources

either because it is not applicable, or because insufficient data was recorded to complete it with confidence.

The variables `source2` and `duration` have missing values exclusively due to non-applicability. As such they have been integrated into (complete) summary variables for analysis, forming the computed words and computer genre in an effort to unify measurement of spoken and written material, and increase the specificity of the data source recorded.

Variable	Missing	Description
source	-	The data source (email, audio, etc.)
source2	34.3%	Any origin data from within the source (i.e. the caller in a phone conversation, sender for email)
day	-	Day of the month sampled
mode	-	Spoken or Written
circulation status	-	Exact circulation, or 'h', 'm', 'l' following Lee's guidelines
portion read	-	The attention index mentioned above
informal genre	-	Free-coded genre
medium	-	The form of the language used. Usually the same as the source.
title	-	The title of the document, or an identifier for the text
author	18.0%	The author[s] of the text
author age	97.2%	The age of the author[s]
author sex	97.1%	The sex of the author[s]
author type	-	An indication of the type of entity authoring the text. Follow's Lee's coding for **s**ingle, **m**ultiple, and **c**ommercial
produced/consumed	-	During the transaction, was text produced, consumed, or engaged in interactively
words	2.8%	Authoritative word count
duration	97.2%	Authoritative duration for spoken data
computed words	-	Estimated word count, merging durations and word counts above
computed genre	-	A composite of the medium and the informal genre, intended as a more specific genre representation

Table 4.1: A summary of variables sampled

4.5.1 Activities

During the sampling period, all work surrounding the personal corpus case study was stopped, in favour of other, ongoing, projects. Other than this intervention, work continued as normal, which for this period involved significant amounts of development on the LWAC download tool (detailed in Chapter 3).

This meant that the bulk of each day was taken up with terminal output events, and editing files. Part of this work involved writing Ruby code, patches from which were used to construct token counts of the 'code/ruby' type. Another significant portion of this time was spent documenting the tool, which required production of longer textual documents (rather than code).

During work, the subject listened either to the radio, or to music, accounting for the large volume of broadcast media and music in the corpus.

Mornings and evenings also usually featured the radio, and often some form of TV (either streamed online or broadcast).

4.5.2 Distribution

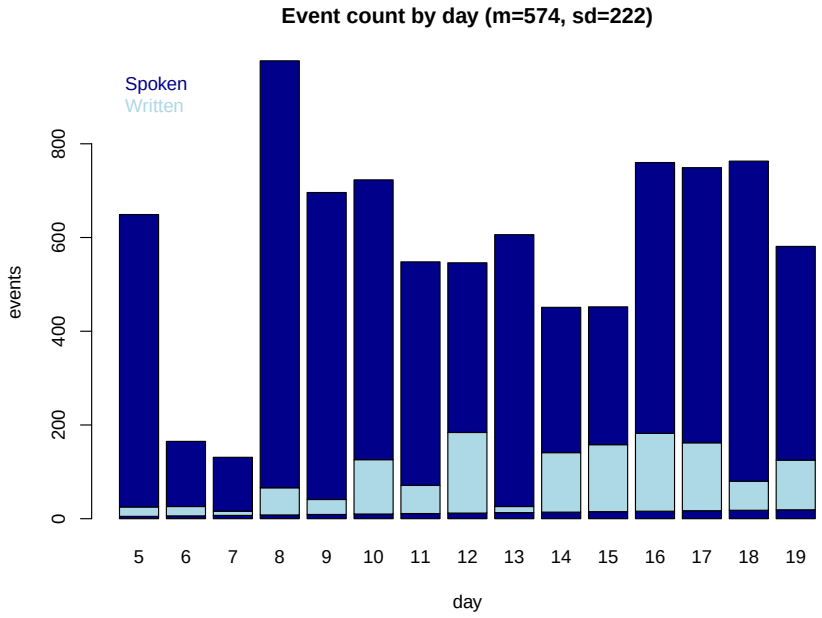


Figure 4.7: Daily event frequencies

Figure 4.7 shows the number of events recorded each day. Event counts ranging from 124 to 969, with much of the variation being expressed in terms of written data.

Contrary to expectations, and in part due to large amounts of terminal and web use during software development work, the majority of daily language use is written. This is contrasted by Figure 4.8, which indicates that spoken events are likely to expunge greater amounts of words in a single event (spoken median is 128 words, vs. 27 words for written). This is largely explainable by a penchant for consuming broadcast media (especially spoken radio), which have large amounts of ongoing spoken content.

As shown in Figure 4.7, the event count as broken down by day shows that between 35,000 and 80,000 words were used each day. As we shall see later, this is mainly explainable by way of broadcast media and web page views. It is notable that weekends fall upon days 6,7,13 and 14. There is no significant pattern by day.

Before this study it was conjectured that the vast majority of all language used would be spoken.

Figure 4.9 shows the breakdown of words used by the subject, by others, or in a manner inseparable for the purposes of the linguistic event (such as conversation). Overall 85% of words were consumed (95% of events), 15% were interactive (just 4% of events) and 0.5% were produced

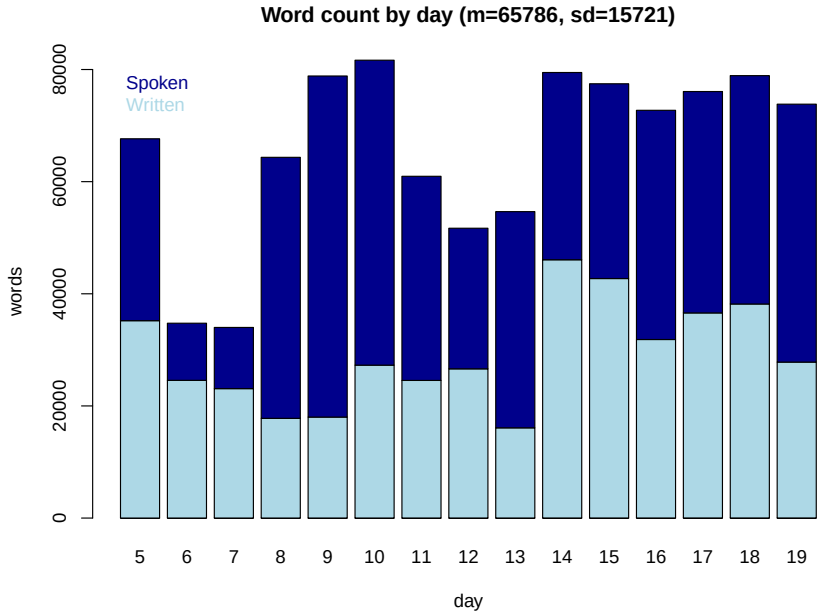


Figure 4.8: Daily word counts

(0.5% of events). This indicates that linguistic events involving some level of interaction yield more words, either by form or duration, something backed up by median frequencies (55 words for production, 27 for consumption, and 80 for interaction events).

4.5.3 Events

The proportions described in Figure 4.10 are likely a very individual trait of the subject, affected not only by daily routine, but also the specific work being undertaken at the time of the sample (the implications of this on scientific validity are discussed below in Section 4.6.6.

Particularly curious is the large number of terminal, web (squid), and music events. These are partially explainable by the work being completed at the time of the study, as the subject was developing software, listening to music whilst doing so and using many web references as documentation. The subject also uses a GUI on his computer that mandates use of terminals for many tasks (this is doubtless atypical of many populations).

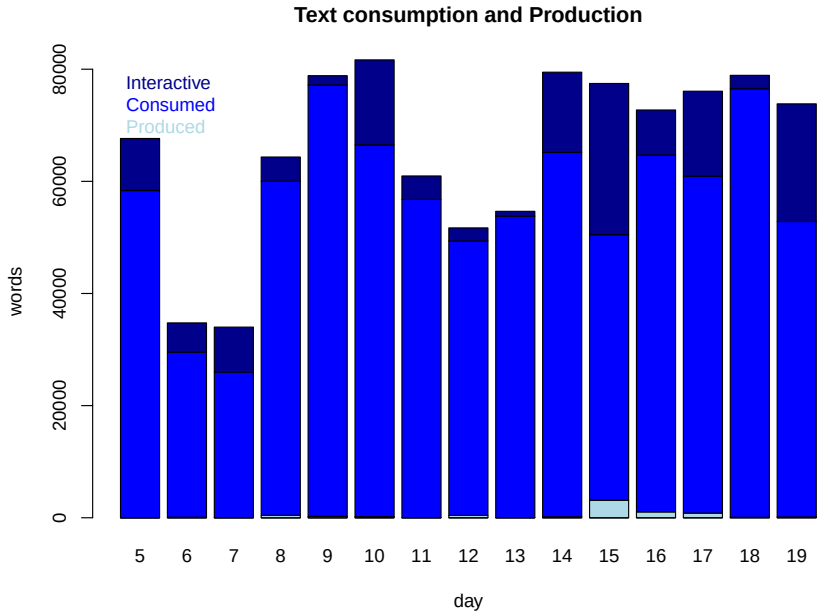


Figure 4.9: Daily word counts broken down by receipient status

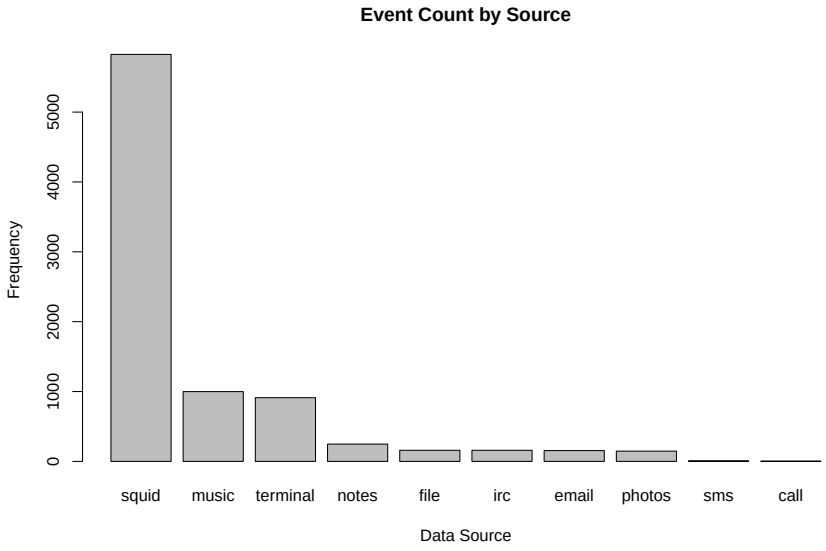
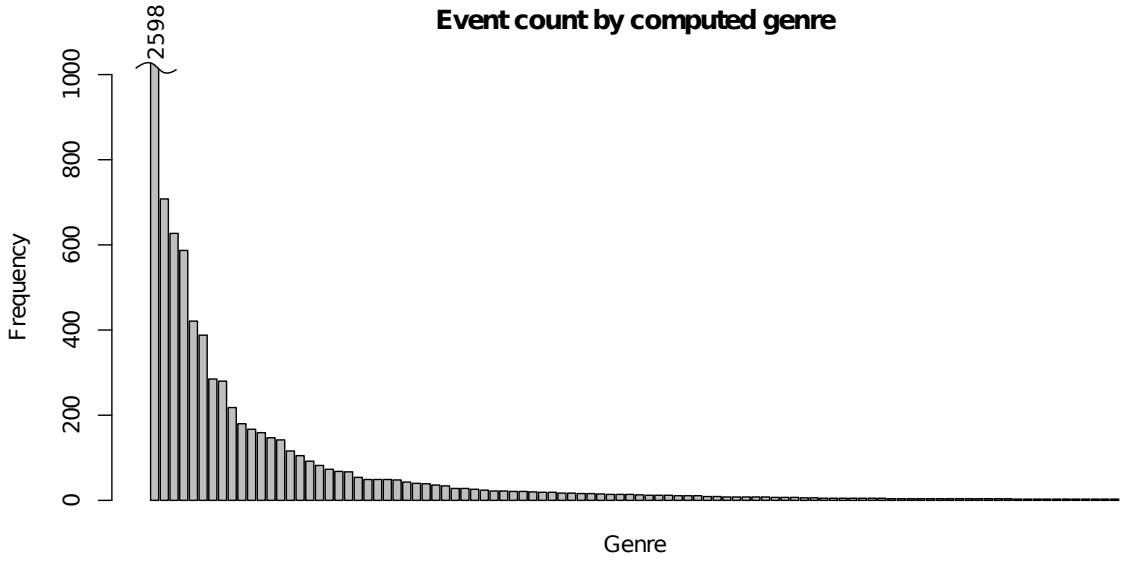


Figure 4.10: Event count by source



Computed Genre	Events	Computed Genre	Events
web/entertainment/social	2598	speech/informal	21
web/music	708	web/comedy/social	21
display/output	627	email/orders	20
music/punk	587	speech/service	19
web/reference	421	web/search	19
web/shopping	388	email/spam	17
display/interactive	285	music/blues	17
web/academic	280	speech/greeting	16
web/tech	218	web/fitness	16
web/outdoors	180	web/tech/news	15
web/video streaming	167	speech/info	14
irc/informal/discussion	159	web/tech/shopping	14
music/guitar	147	web/reference/academic	13
web/academic/reference	142	TV/news	12
web/tech/reference	116	radio/music	12
web/news	105	radio/news	12
file/ruby code	82	sign/info	11
music/metal	73	web/art	11
music/pop	68	web/music/shopping	11
music/rock	67	music/folk/metal	9
web/hosting	54	music/progressive	9
file/documentation	49	email/technical	8
web/comedy	49	sign/info/instruction	8
web/social	49	speech/ephemera	8
web/product support	48	web/banking	8
product/packaging	43	music/country/rock	7
web/tech/academic	40	music/fusion	7
TV/comedy	39	sign/product info	7
web/outdoors/reference	36	book/info	6
email/academic	34	sign/advertising	6
email/advertising	28	TV/documentary	5
web/entertainment	28	flyer/advertising	5
email/personal	26	notes/note	5
file/config	24	product/info	5
email/business	22	speech/parting	5
speech/discussion	22	web/advertising	5
TV/entertainment	22	web/conspiracy	5
		items < 5	180

Table 4.2: Event frequency by computed genre

Table 4.2 shows a breakdown of event frequencies by computed genre (which includes the source). This shows a somewhat predictable Zipfian distribution of genres surrounding a number of themes:

- Daily events, such as reading social media websites;
- Short events, such as consuming single programs of media (especially songs due to additive effects with the above);
- Any of the above that are particularly surrounding personal interests of the subject.

Clearly, the last of these is the most obvious and desirable effect of building a corpus using this method. The overwhelming prevalence of `web/entertainment/social` events may be explained by one website, `imgur.com`, which displays a single image per page (along with socially-contributed captions), and thus demands a lot of page reloads for relatively little content.

The mean words-per-event for each genre is listed in Table 4.5.3. The embedded graph shows a log-distribution of the word count for each, displaying a typical Zipfian distribution but with an unusually fat tail (that drops off sharply at the 5 word level). This would seem to indicate the minimum word length necessary to convey information without presupposing context, (i.e. branding, packaging and signage).

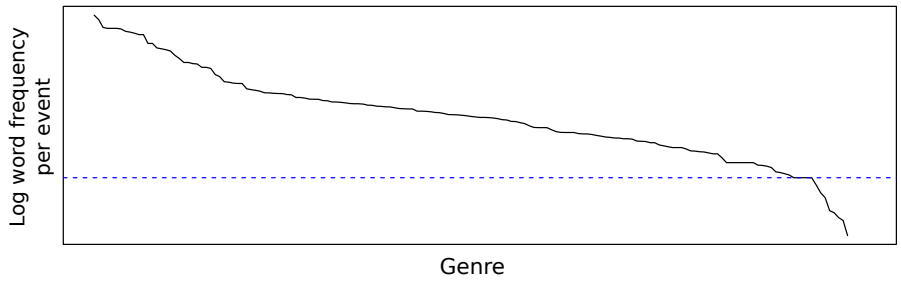
4.5.4 Word Counts

Word counts were established either directly, or by assuming an 80 word-per-minute speech rate. This rate is a conservative estimate according to some literature—Yuan et. al.[187] estimate almost twice this, however, they do so for continuous conversation, and much of the corpus' spoken data comes from media with interruptions such as TV or music. The music included in the corpus typically has speech rates falling between 40 and 70 WPM, a figure easy to calculate using song lengths and lyric sheets.

Table 4.4 shows that the distribution of word count per event is heavily skewed towards the lower end, with the median being 27 words per event.

This figure also reflects the degree to which the attention model affects results—disabling calculations for this, the median word count is affected significantly by the largest source of data, web pages, becoming 442 words. Most pages are, even compared to other types of event in the corpus, read only partially (often graphical clues indicate which portions to read, especially since most people revisit websites more than they visit new ones).

Figure 4.9 shows that the majority of all words are consumed, either passively as the result of receiving broadcast media or in part of a conversation. The vast majority of the 'produced' word count is covered by writing documentation for LWAC (a tool detailed in Chapter 3), and it would seem likely that most people would non-interactively use spoken words.



Genre	Mean Words	Genre	Mean Words
radio/news	4183	web/music/shopping	120
book/info	1950	music/metal	119
TV/news	1876	music/guitar	118
radio/music	1693	file/ruby code	117
speech/informal	1390	web/reference/academic	103
TV/Entertainment	1000	web/outdoors/reference	100
TV/comedy	949	web/academic/reference	95
speech/discussion	930	web/shopping	91
TV/entertainment	760	web/reference	90
email/personal	575	speech/ephemera	90
speech/service	381	music/punk	86
web/social	286	web/entertainment	83
web/product support	278	web/tech/shopping	81
web/comedy/social	266	web/outdoors	77
music/blues	248	file/config	75
email/academic	247	music/fusion	71
file/documentation	238	web/banking	54
speech/info	225	web/academic	50
web/tech/news	194	web/fitness	45
web/tech/reference	185	web/hosting	36
web/news	184	email/technical	33
music/folk/metal	183	display/output	32
web/music	173	web/entertainment/social	26
music/progressive	156	email/spam	17
speech/greeting	150	display/interactive	16
email/advertising	149	web/video streaming	16
music/pop	147	web/search	15
email/orders	140	sign/product info	10
web/art	139	product/packaging	8
irc/informal/discussion	132	sign/info/instruction	8
music/rock	130	sign/info	6
email/business	129	sign/advertising	6
web/tech	124	web/comedy	3

Table 4.3: Log word count by genre and table of mean word count per genre (for all 66 genres with over 5 occurrences)

	5%	10%	20%	50%	70%	90%	95%
1	0.00	0.00	8.90	27.85	74.00	217.00	354.88

Table 4.4: Percentiles for words-per-event.

4.5.5 Comparisons

At a large scale, there are some obvious differences between the personal corpus and the BNC's selection.

One of the most striking of these is the rate of broadcast media consumption, which forms a full 50% of the subject's spoken word count, yet only 10% of the BNC's. There is a fair argument to say that many people of a similar age, who grew up surrounded by online streaming services and easy access to cheap receiving devices, will share this disparity.

A less compelling difference is exhibited between the rate of spoken conversations—our subject's spoken data was 20% conversational, whereas the BNC contains 40% conversational data (as a proportion of its spoken component). The degree to which this applies to others is difficult to estimate, since (contrary to the verbiage of this thesis) the subject may simply be unusually laconic.

Arguments for the diversity of texts within corpora are driven only vaguely by the quantitative findings of this study due to its restricted inter-person relevance. Nonetheless, some trends were identified which could reasonably be expected to extend to many others in a similar culture.

Inspecting the selection of genres yields a diversity not covered in many corpora. Though notably the BNC is quite good at this, covering many brochures and other incidental texts such as unpublished letters or school essays, it hardly represents the amount of times thanks were exchanged for holding a door open, bottles, cans and other labels were read (particularly addresses on mail), or a tax disc was checked. It would seem that a corpus purporting to cover a large population would be awash with weirder texts, especially as those less academic than the subject here are likely to have relatively higher proportions of them in their corpora.

A particularly compelling argument is made for the inclusion of song lyrics. These are a large portion of the corpus collected here (163, 288 words), and often spawn sayings, memes and linguistic affectations that can last decades in popular culture (*"will you do the Fandango?"*). It's also notable that this is one of the easier things to sample and music use by the population is meticulously documented and studied by the industry already (often in publicly available lists). In the case of music, use of English as a lingua franca would see its influence spread far beyond the bounds of a national corpus (as well as absorbing influences from other national flavours of the language).

Genres Absent from the Corpora

There are a number of genres absent from the corpus that are listed in the BNC. Examples of this that are unlikely to change if the sampling period is extended indicate lifestyle differences and, in sum, indicate the likely representativeness of the BNC relative to the subject. Where a genre is seen as likely to be represented given sufficient time, they are evidence that the sampling period was insufficient (something that the BNC's breadth mitigates).

Comparisons were performed between the list of composite source/genres given in Table 4.4 and Lee's genre listings. Where any overlap was judged to have occurred, a genre was discounted entirely.

Medium	Genre Category
Spoken (S_)	classroom courtroom demonstratn interview* lect* sermon
Written (W_)	humanities_arts ac* (all but tech_engin) biography essay* fict_poetry hansard newsp_*_sport pop_lore religion

Table 4.5: Major categories missing from the personal corpus but present in the BNC.

The 11 (of 24) top-level categories missing entirely are summarised in Table 4.5. Of those spoken, only interviews and lectures are likely to ever be encountered by the subject during a longer sampling period. The former of these would require a sampling period of years for most subjects.

Written material is more comprehensively covered, with 15 of 46 genres missing. Whilst newspaper articles are absent from the corpus in their original form, many of the same genre were read online (with the exception of sports articles).

As expected, the genres identified within the personal corpus are an imperfect subset of those featured in the BNC. Following Lee's index, there are no magazines, religious texts, newspapers, dramatic scripts, school or university essays, texts on the subject of commerce/economics or academic texts in many fields. Newspapers alone comprise 9.9% of the BNC, implying that caution should be exercised before applying any findings from the BNC to any single subject.

Any comparison should also take into account the size differential—the BNC is roughly 100 times the size, and it is reasonable to presume that many of these genres would be covered by a 100-person replication of the two-week sampling procedure. The strength of this assumption is something the current experimental design is incapable of evaluating.

Demographic Changes

A number of these genre differences are specifically down to the demographic coverage of the corpus (or lack thereof). Though these differences are to be expected due to differences in corpus design, they illustrate the fallacy of applying results obtained for a population to any individual therein without qualification.

These categories have been selected because they are not only missing from the personal corpus, but are likely to remain missing for the subject, regardless of the sampling duration used.

Such missing categories include:

- School work (and thus many essays);

- Anything relating to religion (sermons, religious texts, etc. are missing entirely, yet are among the most popular books for the wider population);
- Business-oriented documents (due to status as a PhD student);
- Medical texts;
- Many academic subjects' materials.

Temporal Changes

The BNC is now ageing, and some changes are expected that likely generalise to enough of the BNC's population to make a significant difference to its representativeness. Clearly, the personal corpus described here does not rigorously explore these, rather it is a heuristic for how such changes may affect the representativeness of a larger corpus in real-world terms.

Since the BNC pre-dates ubiquitous technology such as smartphones (and, largely, the web), it clearly omits what has been shown to be the largest single source of data in the personal corpus. The severity of this impact is doubtless effected by the technical experience/interests of the subject, who is both relatively young and technically capable, compared to the wider population, however, this trend is widely acknowledged and set to continue.

Arguable is the extent to which other genres have been co-opted by the web, suggesting that it is being used more as a transmission medium than a conventional genre. The depth of this analysis is defined by one's purpose, though it is notable that the breakdown of text genres from web sources in the personal corpus extends Lee's 'email' category into 8 sub-categories, all of which overlap significantly with those used elsewhere. This pattern is repeated for the web (the personal corpus' largest source of data), which yields 68 genres with over 5 entries[25].

Some genres have arisen that do not so easily map onto the original BNC design. Email spam is one, presumably contained within the 'email' category. Another is the use of social networks, which has no direct correspondence in the BNC genre listing, and could be thought of as a form of very informal news, or very informal social interaction in line with public debate (pub_debate) or conversation (conv).

It is unfortunate that the BNC, even with Lee's annotation, lacks sufficient demographic detail to identify where the subject falls within the BNC sampling frame. This would allow approximate weights to be applied in order to rebalance the BNC as an augmentation of the personal corpus.

Particularly interesting are instances where the hierarchical categorisation of the BNC indicates that both corpora have sampled from the same sources, yet the resulting within-category proportions differ. This is true of many of the news sources, where there was a strong bias towards tech. news, and academic sources, which were focused on a single area. We might reasonably expect this effect to be true for many sub-populations, who may prefer a given religion or set of recreational activities.

[25] I have excluded genres with low counts as these are unlikely to deserve their own category in Lee's classification, though a suitably scaled general purpose corpus is likely to include many.

4.6 Discussion & Reflection

The quantitative data presented above are of limited utility to others, since the degree to which they describe the subject's life (rather than some general linguistic trend) is unknown. Without further sampling, or detailed linguistic auxiliary data, we cannot begin to generalise from the above in a useful manner.

This does not mean, however, that we cannot reason rationally about how transferrable those results are—it is more unlikely that a member of the general population is a software developer than it is that they watch TV, or listen to BBC Radio 4 in the mornings.

Moreover, since the purpose of the study is to assess the viability of methods, these may be seen as significantly more transferrable than the data they have collected (this distinction is blurred by the inclusion of a human interest model, however.)

4.6.1 Types of data

The form each corpus takes is rather variable due to the different sampling methods used. For some sources, e.g. books, the personal corpus is likely to contain more partial samples (as books are seldom read in entirety within one 'linguistic event'). Those sources that are read piecemeal are also likely to occur multiple times, something that would not be tolerated in conventional corpus designs.

4.6.2 Method

The process of gathering data itself was optimised for low intrusiveness, and largely proved practical to pursue medium-long term.

Maintenance of the live notebook was the major interruption in everyday language use, and this was gradually optimised to a shorthand format that demanded less time in the field, yet more annotation to extract data. Where the subject is someone other than the analyst, the format of journals would need to be solidified ahead of time to ensure all properties are identifiable.

On reflection, the mnemonic effects of the notebook were most useful in operationalising the data—the narrative they created placed all other data items in context, and the detailed notes in the nightly journal contribute to an episodic memory that aided the coding process. Again, this process of reflection is unlikely to be made available to an analyst without explicit insertion of an interview phase where the two may meet to resolve possible misconceptions.

The nightly journal, whilst useful during the study as a place to write down easily-forgotten details of language events, was less useful than anticipated during analysis. In reality, the richness of the narrative one quickly notes down of a night proved to be difficult to operationalise, meaning that it was largely of use only as auxiliary data to augment the live notebook's event-by-event format.

Some sources of data proved significantly easier to normalise into a workable format than did others.

SQUID logs proved particularly difficult to process largely due to technical reasons, as extensive whitelists and multiple manual inspection phases were necessary to disregard advertising and AJAX requests. A possible solution to this would be to rely on sampling of web data closer to where it is consumed, for example through the use of a browser plugin.

Initial plans were to use Voice Activity Detection (or full automated transcription) to estimate word counts from verbatim audio recordings. In practice the diversity of unwanted noise effects (background noise, positioning of microphone, overheard conversations) rendered this practically impossible. Even rudimentary VAD algorithms work with frequency detection strategies, and many were liable to detect things such as music as continuous speech.

Of particular interest to the methodology is the inclusion of a human interest model in the processing stages. Without this model, the data is changed to massively overestimate the word counts of many data sources (some more than others), and it is the opinion of the subject that the model improves the plausibility of data. Generalising the method requires generalising this model or narrowing down the number of sources it must cover, either by using more selective data gathering methods or by restricting the scope of the sample.

Each of the data sources mentioned is burdened with its own empirical concerns that must be considered when designing a human interest model, and the general solutions for many of these may be particularly complex. For example, models for websites using large graphical elements as well as text must take into account Gestalt principles as well as models for humans as they read text, and the particular interests of the subject.

One way this problem may be solved is through detailed elucidation of particular features and interests via a questionnaire or laboratory tasks in addition to the data gathering methods described here. Clearly, though, relating these to a usefully-complex model of human interest will demand further research.

One of the larger challenges in gathering data from so many sources (and in so many forms) is the amount of work needed to operationalise it. This was done both manually (in the case of complex data such as photographs and notebook entries) and automatically (for the majority of sources). Often, some degree of manual correction or annotation was necessary even where automated processing was used. These processing tools were bespoke, and would need to be generalised if the method is to become viable for use gathering further samples. The tooling used is summarised in Appendix B.

A number of questions were raised during the sampling period surrounding the limits of data that should be captured. The solutions used were taken from the subject's own judgement of what constituted language 'use' (since this case study is predicated upon recording that opinion), however, further work would be needed to establish answers to these in the general case, and other researchers may be guided by more principles linguistic theory.

The attention models used to narrow down input have already been mentioned, however, they apply at a very specific level—often, shorter texts such as signs, brands and labels would have particularly familiar or conventional text placements, styles, and shapes (some media, such as road signs, deliberately accentuate these features to aid recognition). It is unclear at what point

a text is 'read' rather than just recognised in the periphery of one's vision. The solution used in this study was one of internal vocalisation—if the text was recognised sufficient to repeat/act upon it mentally, it was classed as read. This means it is often possible to say that a text source had just a few words read, when in fact a number of boilerplate features were skipped over because of their position and style.

An extension of this problem is one of re-reading texts that are being written. The degree to which this occurs is likely extremely variable by person and task, however, it proves to be a particularly complex form of the above problem and is particularly hard to measure (or even subjectively assess). In this case study no attempt was made to compensate for this effect. Detailed study using some kind of attention measuring system (for example, eye tracking) may allow for deliberately using attention coefficients greater than one (for word counts) or deliberate repetition of data in the final corpus.

The suitability of the attention system used in this case study is also debatable. The current method of using a coefficient works only for estimating word counts, and does not favour certain regions of the final data over others. Additionally, it conflates two effects—that of reading only a small proportion of the source text, and that of paying little attention to the text (you may also consider using a second text at the same time a third form of inattention). Annotation of simultaneous events was fairly simple in the notebooks, however, estimation of the distribution of one's attention was done only at a low resolution (the codes used practically equate to ratios of 80/20, 50/50, 20/80, 100 with only a few exceptions).

Since the aim of this study was to assess genre proportions, significant amounts of effort were saved by not converting all sources into verbatim text. This is largely an issue of post-processing, as many methods were digital and thus yielded verbatim text with ease and those that do not have well established transcription procedures. Of particular note is the importance of being able to apply a regional attention model to extract (or weight) the most read portions of a text, something that (as noted above) was not attempted here due to a focus on proportions and word counts.

4.6.3 Sampling Period & Validity

It seems clear both from personal reflection and examination of the data, which strongly favours certain work-related activities over others, that a two-week sampling period is hardly enough to represent even an individual's language use.

There are a number of obvious reasons for this, particularly egregious examples being:

- Some language sources are usually read slowly, such as books read at bedtime. A sampling period that covers an individual's various literary interests would have to be very long indeed.

- Life is strongly periodic, and though attempts were made to cover the weekly cycle of work and weekend, many events occur annually (either due to seasons or social convention). There are no Christmas songs in my corpus.

- Many more obscure items were represented only once in the corpus (such as advertising on vehicles or documentation on things such as legal forms). These features may not be periodic but they are temporally rare.

- A person's behaviour is likely to be 'chunky', following one pattern for a time and then deviating suddenly (for example, during holidays or upon a job change). Significant effort would have to be given to careful description of duration and circumstance in order to make confident generalisation from the data captured using this method.

It is my opinion that this method is up against two major challenges which must be balanced in order to achieve some degree of scientific validity. The former, and most addressable, of these is the difficulty of sampling in-the-field. The more obtrusive a method, the higher quality data is going to be, however, the shorter any practical sampling period is likely to be. The latter, as already discussed, is the problem of generalising between people in order to create a corpus of inter-person language use that retains the same details discussed here.

Given the relative paucity of metadata within many general-purpose corpora and the periodic/temporal complexity of life, it seems reasonable to suggest that an ideal mix would be better formed from long-term (multi-year) samples using low-resolution sampling techniques. This may entail, for example, discarding the on-line journals and manual photography in favour of automated life-log photography and a nightly journal only.

Following sufficient repetition of short-duration studies, it would be possible to be more confident in the results of longer-term, lower-resolution samples, as the areas with less uncertainty may be attended to less intrusively.

4.6.4 Data and Comparisons

As mentioned in the quantitative section above, there are some large differences between the corpus gathered here and the BNC (here used as an example general-purpose corpus that purports to cover the subject). It is clear that a number of these are down to individual variation (the technical and musical genre bias), however, other disparities are of an altogether more ambiguous status.

The overall word count of the BNC, and other similar corpora, is called into question by the size of this corpus. In two weeks, a single person used almost a million words whilst focusing heavily on just a few subjects and roles. Given the sheer number of people within the British population, their demographic variability, and the dubious extent to which even these two weeks represents our subject's use of language accurately, it seems preposterous to suggest that a mere 100 million (or even billion) words would sufficiently represent language use for almost any purpose.

As previously discussed, the different selection of genres indicates that the BNC's selection of materials is an imperfect superset of the subject's use of language. It is reassuring that many of the genres absent from the BNC are minor ephemera (signs, tax discs etc.), however, there is a significant underrepresentation of others (mainly technical and broadcast media). This

underrepresentation is partially due to the subject's demographics, something that we would expect to skew the relative proportions of the corpus, and partially due to the age of the BNC, something that is particularly undesirable.

4.6.5 Validity

The design of this experiment is subject to a number of challenges to validity.

Perhaps the most severe of these is the extremely personal nature of the corpus itself, which renders verification of the data all but impossible except through elicitation and subjective judgement. This is to some degree a property of all case studies, especially those seeking to experiment with methodology.

A strong case has been made in many fields (and by the inaccuracy of certain assumptions made during preliminary tests within this study) for the fallibility of subjective opinion, and this is partially the reasoning behind a census methodology—it is a simpler (and hopefully less controversial) task to mechanically and objectively record each linguistic event than it is to estimate their size.

The main subjective component of the method used here lies before the linguistic events themselves are recorded: judging when a text is read, rather than unthinkingly 'seen'. The assumption made for this case study is hopefully uncontroversial enough to be accepted for a majority of purposes: after all, it seems unlikely that we will be able to develop an objective and meaningful threshold for this.

The primary challenge to validity due to subjective reasoning occurs after the data is captured. This is where issues that lie beyond the scope of this thesis are—development of a robust and generalisable human interest model being a major one that has been shown to make a large difference to the results. The model used for this study is deliberately and knowingly subjective, and would not be applicable to any replication effort.

From another perspective, the data set described here is difficult to relate to existing literature due to its orthogonal sampling structure—the whole corpus represents a single (albeit very rich) data point in most other corpus designs, and this robs us of quantitative knowledge of how it relates to other data sources.

This question of generalisability has been attempted by a rational comparison with corpora of known demographic coverage above. A better method still would be to extract only those texts from a corpus that match the demographics of the subject described here. Unfortunately, this is not possible to any useful degree given the limited information on the users of texts within conventional corpora, and so a more detailed comparison is again stopped by the lack of known-similar data.

These limits on comparison to others are both lifted by restriction of the data types being covered, especially where those types are easy to sample. It would be possible, for example, to build a special-purpose web history corpus in which to contextualise a user's web history, and use this to impute their position in larger corpora with greater-than-zero confidence.

4.6.6 Ethics

The increased resolution of data pertaining to a single individual renders the methods discussed here ethically sensitive. This sensitivity is increased further if continuous recording of audio or video are used, though, as mentioned above, this data was not integrated into my analysis.

Future developments in the methods described may use questionnaires or other less-invasive methods as sources of auxiliary data. These would be targeted to a particular study design and need not cover the full set of language uses, mitigating any ethical concerns by limiting the descriptive power of the raw data itself.

A number of technical measures are also possible that may assist this issue—some of these have been developed by Ellis & Lee working on the Machine Listening project[43], who irreversibly scramble their audio recordings in such a way that VAD algorithms may still run. A further option is streaming of data to a remote server, which can process, summarise, and discard verbatim data on-the-fly to prevent any possible information security breaches.

Particular sources of data raise various ethical and legal issues. These are generally centred around the presumption of privacy, which applies to most direct personal interaction, as well as some more formal communications (such as email)[144]. This relationship is described by Herrera[73]:

> *Covert research is also not, on reflection, so much like the conversations that we casually engage in. On the contrary, covert research raises a unique problem: those closest and most vital to the 'conversation' are valuable only so long as they know less about it than anyone else.*

Many of the ethical objections posited in sociological texts apply to the misrepresentation of a field worker, leading to an asymmetric relationship between him and other participants, and necessitating various abrogations of personal and professional ethics. These issues already affect life-logging for personal reasons, something that is differentiated largely by how widely the data is disseminated.

Any use of covert methods should always be justified in terms of the importance of the research, lack of alternatives, and sensitivity of the data to be gathered. Clearly, recording audio will require a higher standard of justification upon each of these, and the counterarguments below are written with both approximate (recording genre, setting etc.) and verbatim methods in mind:

- Eliciting consent would disrupt linguistic content, something that is minor for large interactions but makes representation of shorter utterances and conversations impossible. As one aim of the sampling design used here is to capture a more empirical dataset, any attempts to elicit consent must be less disruptive than those used to build other corpora.
- Without survey of all language used, it is impossible to assess the validity of hypotheses in the study.

- Language data recorded is unlikely to be directly personally identifiable, or culturally/-commercially/governmentally sensitive. This will depend on the status and role of the researcher.
- Results may be reported without release of the data to third parties. The lack of generalisability from having a single data point means that there is little value in releasing the data anyway.
- The study does not centre around sociological issues, and my participation in events will not be subject to judgement or inquiry. Simply, social issues are tangental to the content of the study (though it is notable that some readers may find more interest in them than the author intended).

The BNC, when building its speech corpus, followed a procedure whereby consent was sought but only after each conversation[31, p. 21]:

> *All conversations were recorded as unobtrusively as possible, so that the material gathered approximated closely to natural, spontaneous speech. In many cases the only person aware that the conversation was being taped was the person carrying the recorder.*
>
> *For each conversational exchange the person carrying the recorder told all participants they had been recorded and explained why. Whenever possible this happened after the conversation had taken place. If any participant was unhappy about being recorded the recording was erased. During the project around 700 hours of recordings were gathered.*

Though this is a higher standard of consent than sought for this study, we feel this is offset by the fact that our data is private and reviewed only by the subject himself. This is a justification reinforced by the Human Rights Act (1998), which guarantees privacy to prevent release of unsolicited recordings, but makes no restriction on the use of such data by a participant in the original conversation.

4.6.7 Future Work

In the long term, it is hoped that a greater understanding of the above may contribute to:

- Methods for augmenting and rebalancing corpora by identifying the position of a subject within a larger corpus' population;
- A greater understanding of variance in terms of the populations being studied

From a sample of just one person, it is possible to use auxiliary data from existing sources to operationalise and reason about inter-person variability. This may be done by cross-referencing a subject's demographic variables with those from an existing corpus, placing them in context and allowing comparison of his linguistic data to other groups (or to those within a given similarity). This technique can also be used to impute data from partially-sampled sources, creating a personal corpus by re-weighting existing samples.

Unfortunately many existing corpora are unsuitable for this process due to the limited availability of metadata (something that is also an issue for those constructing 'informal' subsets).

Methodologically, it is possible to generalise the data-gathering procedures mentioned in this case study either through reduction of the population covered (the technique used here in extremis), or reduction in the linguistic fields covered (the technique used more typically in special-purpose corpora). A combination of these two approaches may well lead to the technique being used for many questions.

Careful application of elicitation strategies such as questionnaires or source-specific tools like web usage monitors may be able to produce sufficient auxiliary data to resample larger corpora, however, these methods would need justification from repeated studies such as the one described here.

4.7 Summary

It is clear that there are significant differences between the language experienced by the subject here and the proportions of any BNC-like corpus. These differences are partially explainable by temporal effects (such as the rise of the web and portable availability of broadcast media) and partially by demographic. Those non-demographic differences imply that the BNC would apply poorly to the population as a whole, if used as a sample of language exposure.

There are a number of methodological challenges that continue to prevent application of the methods described here to larger populations. These are either due to the problems inherent in sampling multi-modal data in the first place (digitising photographs or transcribing audio), or the processing required to transform data into a usable form (models of attention).

It is hoped that further work will be able to develop these methods in order to mitigate these issues, at least for certain applications. However, the utility of these methods is retained under less ambitious conditions of restricting a combination of the domain (for example, only sampling web histories or work-day language use) and the population (i.e. only people very similar to myself).

The practical issues encountered during sampling are largely minor, and modern portable technology proved decisive in making capture of hitherto-unseen (or at least widely ignored) sources of text possible in-the-field. The census design lends an empirical justification to inclusion of these genres in larger general-purpose corpora, however, we are unable to formally generalise with confidence to demographics other than that of the subject covered.

Of particular note, and perhaps greatest generalisability, is the number of words used throughout the sampling period. Almost a million words were input and output by a single individual over a two week period. Since it is reasonable to deduce that this sampling is at best representative only of a normal working week for a single individual, it is rational to suppose that a corpus purporting to cover the whole population of a country demands a sample size far greater than many existing corpora.

This would seem to form an argument that it is simply not cost-effective to build a conventional

general-purpose corpus large enough to represent large populations, suggesting that the focus should lie in those built for more specialist purposes, or those sampled non-probabilistically. Here, the extreme size of web corpora may be well justified, so long as they can be shown to have been sampled rigorously.

This suggestion is reinforced by the obvious variability in the proportions of texts taken from each source, which will vary greatly by lifestyle. Indeed, it seems that the measure of one's lifestyle by text source is a particularly direct way of characterising inter-person variability, and further work may be focused on this problem as a way of simplifying the methods described here (for example, through the use of cluster sampling).

Of interest to this thesis is the high level of detail that may be captured using this method. The census design affords full coverage of a given area about which we may wish to generalise, and the detail it is possible to capture this way makes the data ideal as a 'seed' for constructing larger corpora with comparable properties. The complementary nature of the sample design here offers a way to mitigate a number of problems surrounding identification of metadata distributions in larger corpora, something that will ultimately increase corpus quality overall.

Chapter 5

Describing, Building and Rebuilding

We have seen in previous chapters, a number of scientific and practical issues that affect current corpus collection methods. Many of these surround stratification and selection of documents for corpora, particularly when applied to specific research questions.

This has typically been a problem for those building 'general purpose' corpora due to the large variety of resulting studies. Johansson, in the LOB manual[87], notes:

> *The text must be coded in such a way that it can be used maximally efficiently in linguistic research. As the possible uses are many and difficult to foresee, the main guiding principle has been to produce a faithful representation of the text with as little loss of information as possible.*

This chapter details the design and implementation of a method for performing a number of common corpus building workflows which allow per-use adaptation of these coding stages. These workflows allow for application of stratified sample designs whilst working with existing concepts in corpus linguistics, allowing for replication and progression of current corpus designs.

The chapter begins in Section 5.1 with an outline of the workflows involved, detailing the rationale behind each, and the value that formalising the process may yield. Immediately following in Section 5.3 is an outline of the design of the method, specifying the processes involved and comparing them both to existing procedures and to the original workflows.

Section 5.4 describes the mechanisms used to implement these designs, and covers the architecture of the software used to test the method. It discusses the algorithms used, as well as the variables selected for the test system and the strengths and weaknesses of these as a mechanism for evaluating the method described earlier. This section then specifies the technical details of the implementation, and provides details on the approximation and optimisation methods used.

Finally, Section 5.5 relates the design and implementation of the method back to the aims detailed in Section 5.1, drawing comparisons again to the capabilities of other existing systems.

5.1 Rationale

Existing methods of study in corpus linguistics centre themselves around re-use of large, known datasets. This, as detailed in Chapters 2–4, is potentially the cause of many biases—replication is unlikely to use a comparable dataset, and many studies test their hypotheses on the same dataset from which they derived the initial idea.

This re-use of data is common to many other fields (such as sociology[129] and computer vision[68]) where raw data is fundamentally difficult to acquire, however, this need not be the case: the internet offers a source of new text with little barrier to retrieval, offering access to a model closer to that used by many other sciences, where datasets are built with a specific research question in mind, and may be replicated from the original population without relying on the same verbatim data.

Corpora built from online sources are nothing new, and they are largely constructed with ease by automated tools that are able to emulate existing corpus proportions.

The 'conventional' method of sampling data is to specify parameters external to the text, such as medium or context, which are (with the exception of the relationship specified in the alternate hypothesis) uncorrelated with its content. The process of relating these specifications to the location of documents to be selected is performed, in lieu of a census of such metadata in the real world, largely through expert opinion. Often this means reliance on partial indices such as bestseller lists or library loan records, or simply the fixing of the proportions used (as used in the demographic portion of the BNC spoken data).

The lack of availability of consistent metadata online has led the authors of web corpus sampling tools to largely use an intensive query system, providing values for linguistic variables (such as example n-grams) and returning 'similar' documents through the use of search engines—either pre-existing consumer ones such as Google or Yahoo, or those specifically-built for linguistic purposes. Where a study is interested in frequencies that are correlated with these terms (something that is often difficult to establish even hypothetically), this approach is statistically fallacious. The value of this approach is in the high dimensionality of the input data—it is deemed unlikely that any one variable of interest is 'truly' correlated.

In order to apply sampling of external variables to the web, it is necessary to algorithmically approximate the process expert corpus builders follow in translating a desired document's properties ('from genre x, of length y') into its location ('genre x is found in libraries, transcribe random extract'). This process is particularly difficult online due to the inconsistent nature of metadata, and the limited scope of indexes available, however, there is great value in moving the assumptions of expert opinion into the sampling tool:

- Reliability—Algorithmic representations can be precisely repeated. Though it may be necessary to change and improve heuristics over time (in much the same manner any body of expert opinion is likely to iteratively improve) it is possible to identify and evaluate these changes to place samples in historical context.

- Documentation—Because of the above, the use of each script becomes a source of docu-

mentation for the corpus, detailing not just what is in it (and the sampling proportions of each) but also the policies used to decide that. This level of documentation allows for comparison against the research question for which the corpus is to be used at a much more useful level.

- Distribution—Codification of expert opinion allows contribution from around the world in an open-source model. This essentially democratises the expert portions of corpus building by parameterising them.

- Repetition—Reducing the human input required to take a sample allows faster repetition, which in turn allows samples with different sampling units. Fast heuristics would allow for a virtual implementation of the library metaphor[46], retrieving a unique document for each word (or any other sub-document sampling unit) and eliminating the need to account for dispersion in corpus analyses.

One of the reasons the parameterisation of expert opinion has not been attempted is because of the high dimensionality of many corpora—it is difficult to usefully specify the contents of a corpus and retrieve them without conforming to a very complex set of interactions. This is a well-known problem in Bayesian statistics, and is often solved by the application of Markov-chain Monte-Carlo (MCMC) techniques[168].

MCMC sampling methods iteratively construct a markov chain of samples from easy-to-determine distributions, the sum of which (and thus the result of the chain) converges towards the desired distribution. This approach can be used to reduce the dimensionality of requests for web data whilst ensuring that the overall corpus still approaches the desired properties.

5.2 Use Cases

The use of external metadata in description of a corpus allows a number of different uses for the method described here:

- **Construction**
  Manually specifying the distributions of each external variable allows the corpus to be built from scratch according to a given sampling policy.

- **Distribution**
  Distribution only of the corpus description document simplifies copyright issues and potentially reduces the amount of data transferred, relying on the end user to reconstruct a corpus and allowing known bounds of variance in the properties to be measured.

- **Rescaling**
  A corpus may be profiled and rebuilt or augmented to resize it whilst retaining the same sampling policy. This is especially valuable where a corpus needs to be augmented with new data, but where there is no reason to discard the original documents.

- **Replication**
  Replication studies are able to use the same corpus description but operate on new data.

- **Repair**
  It is possible to repair a corpus that has missing documents (e.g. due to link rot) by resampling them to augment the corpus until the overall distribution is similar.

- **Anonymisation**
  Distribution of metadata only allows otherwise-sensitive corpora to be disseminated.

5.3 Design

The method proposed here is built around two processes, each of which may be implemented in a number of ways (including through manual processing):

1. Corpus profiling, which produces a metadata-only description of the corpus as a multivariate distribution;

2. Targeted retrieval, attempting to produce a corpus with the same distribution.

The latter of these may be accomplished most easily by a process of bootstrapping: using Monte Carlo methods to sample from a conditional distribution. These samples can then be sought a number of ways, such as manual (or crowdsourced) selection of documents, or use of existing search tools.

5.3.1 Profiling

The process of constructing a corpus description is outlined in Figure 5.1. This may be started either from a seed corpus, or from direct user design.

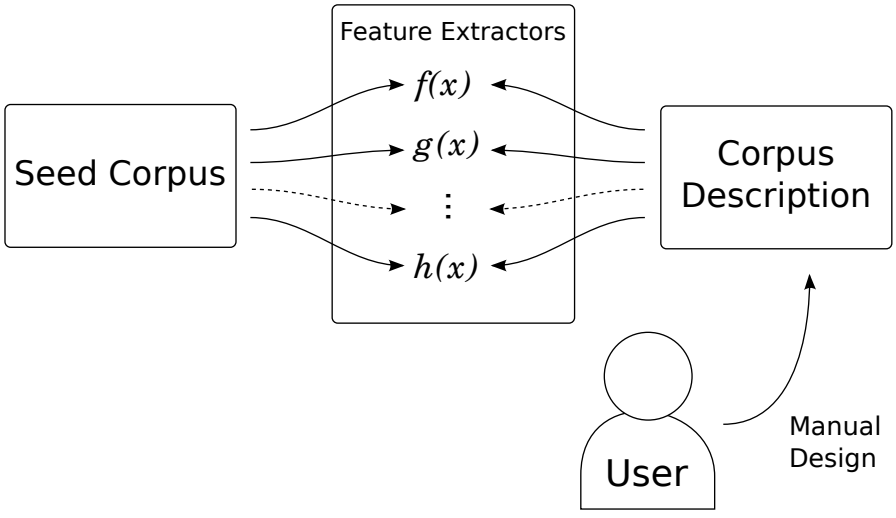


Figure 5.1: Creation of a corpus description

For direct use, the user would specify their salient dimensions, and the distribution for each. This is essentially a description of the sampling policy—where variables are discrete and nominal (such as genre labels), this would take the form of a table with desired proportions against each. Where variables are continuous, a probability density function is defined. Note that this

method does not, without prohibitive complexity, allow for specification of interactions between variables, leaving it unable to represent, for example, differing genres within spoken and written portions of a corpus.

Profiling through the use of a 'seed' corpus begins with specifications of the salient dimensions, the data for which are then read from the seed (by $f()$ and $g()$ functions in the figure). As each document is read, the corpus description may contain information not only on marginal distributions for each variable, but also the effects of conditioning on one or more value.

Note that, whichever method is used, this stage is merely a metadata description task. The corpus description document itself contains no more information than the metadata of the corpus from which is it built—for specifications involving full data on interactions this reduces to a list of points in vector space. For a corpus designed only by specifying marginal distributions, it constitutes a distribution for each. An example profile, compatible with the implementation used in this thesis, is shown later in Listing 5.1.

Those specifying variables to describe must, as with all sampling, be mindful of potential systematic correlations with variables of interest to any given research question. The value of the approach presented here is in documentation—it is possible for a researcher to eyeball the variables (and potentially their distributions) and determine whether or not a corpus is correctly conditioned for use inferring about a given variable. This cannot be said for bootstrapping systems that use internal variables, which seek to copy corpus contents at the risk of varying their metadata.

5.3.2 Retrieval

Retrieval in accordance with the complex distributions specified in a corpus description is challenging in two ways:

- Samples must be taken from the corpus description in accordance with a complex empirically-determined distribution;
- Any combination of variables sampled from the distribution must then be sought based on its metadata alone.

The former of these is difficult because many seed corpora may lack sufficient data to specify their distributions, and because of the high dimensionality of the distributions in question. These issues may be addressed using techniques such as Gibbs, slice, or rejection sampling[57, 133].

The sampling of 'prototype' documents from the metadata distribution is possible regardless of the manner of its construction (however, those specified as marginal distributions will lack any interactions). Here we assume a full specification, since this method is the more comprehensive case.

The following subsections will describe these processes in more detail.

Sampling Metadata

Sampling from the seed corpus' profile produces a 'prototype' document, which has values of metadata fields that, over time, hold the same distribution as those in the seed corpus.

The process of constructing a new corpus is one of continually producing these 'prototypes' conditional on the values of metadata already sampled, and then retrieving texts matching each.

To perform this resampling, we must be able to sample from the distribution of each variable conditional on all of the others, requiring that the corpus description contains information on interactions between values. As such it is only possible if the original corpus description document contains this information, something that is only practically a product of describing an existing set of documents (as manually filling in the values would be prohibitively time-consuming).

Two methods were implemented for selection of the prototype, representing two of the most popular use-cases.

- Simple selection from marginals. This is the best case possible where a corpus lacks interaction information, and follows the joint distribution of the metadata properties across the whole corpus. Despite its relative simplicity it may be suitable for some corpus designs, and is analogous to the BNC's balancing of spoken corpora.
- Full conditional selection. This is implemented by recursively selecting a random sequence of metadata types, and then conditioning on a value drawn from the distribution of that type conditional on the values selected for those previous to it.

It is additionally possible to sample conditioning on fewer variables (yet randomly selecting them each time) in order to emulate the behaviour of a 'blocked' Gibbs' sampler[168].

The former of these, whilst faster and simpler, will fail to take into account any interactions between metadata and is included since it is the only method capable of running on a very simply specified corpus description. The experiments in this thesis use full conditional selection, as they have the luxury of using a corpus description built from a full seed corpus.

Seeking the Prototype Document

The main practical problem of sampling according to the original corpus' distribution is now framed as an information retrieval task—seeking a document that has the same metadata as the one we sampled from seed corpus' profile. Repeating this 'sample and seek' behaviour forms the bootstrapping portion of the method, converging on the input distribution as the output size grows.

This retrieval task is challenging on its own, but errors in performing it have specific ramifications for this sampling process:

- **Dimensionality**
  As the number of metadata properties in the prototype increases, the search space from which a document must be selected increases exponentially[26]. This has severe ramifications

[26] Or, strictly, somewhere between linearly and exponentially, if the metadata properties are not truly orthogonal

for rejection sampling techniques, which become intractable where the ratio of the search space to the area under the target distribution is high.

- **Selection of dimensions**
  Metadata types are selected for research reasons and do not necessarily correspond easily to existing online indexes or retrieval methods. This means that specifically seeking a document with one value of metadata may be a challenging task.

- **Error in selection**
  In part because of the above point, documents retrieved are unlikely to be a perfect match against the prototype. Errors in multivariate selection will follow the population distribution conditional on those metadata values being sought: something that will introduce systematic (yet measurable) bias in the dataset[27].

The retrieval of a prototype, then, is limited to approaches that take into account not only the value of a metadata property to be sought, but also the nature of the source (in this case the internet) and any interactions that are likely to expand the search space to intractable proportions. It should also, of course, minimise error in selection.

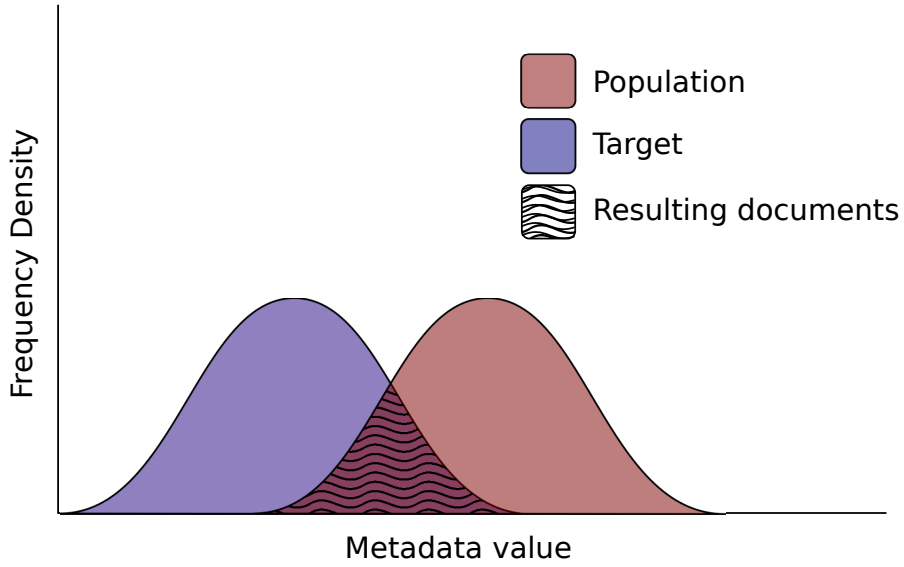


Figure 5.2: The influence of a biased population distribution on metadata selection

In such a case that the retrieved document is a perfect match to the prototype, the output set of documents will converge to the target distribution. If it is imperfect but randomly so, there will be an increase in variance in the output distribution. If it is imperfect and the population is not a superset of the input distribution (or, in the worst case, does not overlap even slightly with the target), then the results will converge to a distribution with bias following the population distribution. In terms of Figure 5.2, we rely on the blue distribution falling entirely within the red one.

It is unlikely that the search space for documents in the population will *always* encompass the range of the input corpus. In such cases the population will bias the output corpus, something

[27] The one, extremely unlikely, exception to this is if the population distribution happens to be the same as the target.

that is also evident in existing bootstrapping tools such as BootCaT[14]. The rejection sampling approach used here, however, makes it possible to measure the difference between the document selected for output and the prototype; providing a measure of error for the sampling method in a manner inaccessible to conventional bootstrapping. This is essentially controlling for population and search engine effects that are otherwise conflated with the distribution of the input data.

Instead of sampling a corpus *of* the web, we are sampling one *from* it, on assumption that the web is already a poorly-indexed supercorpus of documents that are diverse enough to satisfy the majority of research questions. In terms of Figure 5.2, we assume that the distribution in red has a very high variance (not just in one dimension as illustrated, but also in interaction with other metadata). Clearly, the case in which this assumption is least strong is the replication of web corpora, and it is perhaps most strong for speech or other particularly specialist sources such as medical notes.

There are multiple possible approaches to this retrieval task, which mirror the web corpus building approaches of searching and sorting. Each type of metadata may use one or many of these, depending on their relationship to the structure of the web. They are listed here in descending order of 'direct applied expert opinion': the first item in the list relatively dispassionately seeks similarity, and the last uses human judgement directly to identify documents.

i. Searching Using similar methods to BootCaT and other seed-based corpora, it is possible to identify documents conforming to certain metadata values by using a search engine. This encounters minor difficulties in that it is often necessary to transform an ostensive definition (e.g. the domain from which a text is drawn) into an intensional one (i.e. some keywords or other features typifying the domain) in order to fit the format desired by search engines.

If the method detailed here is used only for metadata types which are retrieved by example (the intensional definition above), then this technique for document-seeking is functionally identical to BootCaT's approach.

ii. Directed Crawling Another technique widely used in existing tools is crawling. When downloading documents using a spider, it is common to identify the next set of links (the 'fringe') and then select from the fringe using a suitability metric.

The problem of directing a spider is then one of relating the features visible about a link (link text, url features, position in text, etc.) to the value of a document that will be gathered. This will often involve similar issues to the searching method: it is necessary to examine document content and compare it to some ideal of each metadata value in order to form the link.

Many spiders currently focus on document 'goodness' by simple definitions of information content or URL features[154, 49, 18]—these produce samples that are designed to approach simple random samples (SRS) in order to approximate the population of sites over which they run whilst exploiting the hyperlinked structure of the web. The approach required to deliberately skew this search is arguably an easier problem: a spider may follow rules based on keywords or site features, rather than modelling the content of documents in great detail.

The starting point of each spider may itself be determined by using a search engine. This is common practice, used to direct spiders to relevant starting points, however, the ratio of crawling/searching used for any given selection may greatly change the output.

iii. Existing Indices The use of existing indices for certain types of metadata is a main technique used in conventional corpus construction. Online, however, this avenue is often neglected (with the notable exception of WebCorp[148] and Barbaresi's examination of indices as seed URL sources[11]) despite the existence of many subject-specific indices and general-purpose manually-curated web directories.

By aligning the taxonomies used in metadata to those in an existing web directory, it becomes possible to use the contents of the directory to access documents authoritatively categorised by humans (or to start crawling from these).

This method is subject to errors regarding the age of content on web directories (the popularity of which is waning in favour of search engine technology), as well as those introduced by human categorisers whose aims may not be aligned to the definitions required for a given research question. Additionally, the granularity of each category is fixed and cannot reasonably be changed.

Nonetheless, such an approach is close to that used in conventional corpora and is easily defended, providing that the directory is well trusted. Notably, these directories are already used by search engines to inform their results[167].

iv. Manual Selection Crowdsourcing services offer a cheap mechanism for retrieving documents by requesting that humans manually seek a document fitting the prototype metadata. This method has severe volume and speed limitations not present in the others, however, the accuracy of a human search is likely to be far greater, minimising overdispersion in the output set.

Where a corpus must be duplicated as accurately as possible (yet there is less importance placed on scale or speed of rebuilding), this is most likely a good choice for almost all metadata properties.

The major disadvantage of this method is that the speed of lookup is greatly reduced, bringing the corpus building process closer to that used by a conventional corpus. The main difference here is that the weight of expertise required to select documents is handed off to those constructing the corpus definition, rather than those searching the indices.

Candidate Selection

Since each retrieved document may be missing metadata in some dimensions, it is necessary to infer the values for these from the document itself. This process may use external variables such as HTTP headers/meta-tags in the header of the document, or internal ones that are assessed using some measure unrelated to the variables of interest to those constructing the corpus. This process is performed by a number of heuristics, and may be thought of as an inverse function to the profile generating functions shown in Figure 5.1.

Documents may be ranked by similarity in vector space only after normalising each metadata distance metric (otherwise, distance metrics liable to output large numbers would be apportioned a larger importance in overall fit). This method also allows imposition of weights to specify which dimensions must fit more accurately.

Iterating to Minimise Error

It is necessary to iterate an arbitrary number of times to retrieve a full corpus. If each prototype document is matched closely, a simple repetition of the above will return a corpus that converges to the desired distribution, however, as mentioned in Section 5.3.2, this approach may incorporate bias from the population.

There are two possible approaches to working with biased populations:

- **Feedback from output to target distribution**
  Adjust the target distribution to reduce the probability for those documents already selected.
- **Rejection sampling**
  Select a large number of candidate documents and then identify the document (or subset of documents) that best fits the prototype

The former of these approaches may seem ideal, however, it will only be effective if sufficient documents have already been selected for each conditional distribution, and if the population distribution is only slightly biased. It will also slow down the seeking stage of each document and, in the event that the population yields no perfect documents, stop it completely. It is a stricter and more reliable approach, but also vastly reduces the practicality of any system using it.

Variations on this theme form the basis of many MCMC methods. MCMC algorithms are generally used where it is difficult to estimate marginals, however, this is not the case for our corpus data, where the marginals are well known (or manually specified).

The latter, rejection sampling, is much more viable, as the clustered and hyperlinked nature of the internet makes retrieval of many similar documents only slightly more difficult than retrieval of just one.

The (hyper-)volume of the search space increases greatly as dimensions are added. This is largely mitigated by the directed document hunting approach above, which seeks to minimise the distance between each dimension's sampling distribution and the target. Simply, the better the heuristics are (measured in terms of precision/recall), the more tractable any selection is and the fewer documents that need selecting. The number of documents that needs selecting is proportional to the ratio of the areas under the joint distributions of the PDF of the heuristics' selections over those of the target distribution:

$$numdocs \propto \int \prod heuristics / \int \prod target$$

This presumes that each of the dimensions selected is entirely orthogonal, but serves to illustrate the curse of dimensionality that applies to the search space.

Where a target is defined in terms only of its marginals (as is likely when constructing a corpus from scratch), the interactions observed will be defined by the source population (in this case online), and the sample will exhibit conditional distributions that are peculiar to the internet regardless of its marginal distributions. This approach relaxes the original sampling criteria applied to the seed corpus too, making replication far simpler at the expense of any best-effort approach to minimising bias across multiple dimensions.

As the population distribution for each dimension is unknown (especially if one considers interactions), it is impossible to know to what degree it is being undersampled. This is another source of bias, however, it is one that is clearly labelled as such: those using a heuristic would do well to know its theoretical underpinnings and the methods it uses to select data (just as any corpus user should read the manual).

The main benefit of rejection sampling is that we are able to degrade the performance of the algorithm gracefully. If, having sampled n documents, we wish to relax the conditions of a match with the prototype, this will have only a minor effect on the corpus (increasing its variance and applying a small bias in the direction of the difference between the population and target modes).

The degree to (and conditions under) which this relaxation occurs must be determined according to the particular demands of the application for which a sample is being built.

5.3.3 Measuring Performance

A system built with contingency for accepting certain levels of error is of very little use without some method for measuring it in operationalisable terms. A number of properties of the system are measurable, each yielding information that is useful under different circumstances:

Heuristic Quality The quality of each heuristic may be measured according to its ability to return documents satisfying the one metadata value it is seeking. This may be expressed in terms of precision and recall, or an aggregate thereof such as an F-score.

The choice of evaluation metric will vary according to the type of metadata and its availability online. Heuristics operating in particularly large search spaces (such as those primarily using search engines) will desire a greater weight on precision, whereas those seeking rare (online) data will favour recall. Since heuristics are, ideally, already the results of well-established methods in linguistics, this score and its desirable range should be established empirically prior to use.

The F score may also be generated only for the best documents returned, providing a score for the whole 'batch' of documents returned by a heuristic. This measure is likely to be more representative of stateful retrieval methods (i.e. those that crawl), however, it conflates the statistics used for measuring document similarity with those used to retrieve them.

Ideally, heuristics form pluggable, reliable modules with known degrees of error that may be ignored when using the method as a source of linguistic data. The degree to which this applies is defined by how well each is aligned with a source of data: Keyword-based heuristics, for example, are well aligned with the search systems of search engines, but poorly aligned with the strengths and weaknesses of humans seeking a document type.

The value a heuristic yields is a trade-off between how closely it describes the document and how accurately it is possible to retrieve a document from the source.

Document Distance It is necessary to identify differences between documents' metadata values both in order to select them from the set returned by each heuristic and to identify overall deviance from the target distribution. The former is computed after documents have been retrieved by a search method but prior to their selection; the latter after selection.

Boolean selection criteria (are the metadata in this document equal to the prototype) offer a way to guarantee convergence to the target distribution. They are an ideal case which will be tractable in only certain circumstances.

Relaxing the similarity requirement requires a finer-grained distance metric for each metadata value. The distance for each candidate document may then be measured and a 'best fit' selection made. Alternatively, upper bounds may be set on suitable distances from the prototype for each metadata type, leading to selection of all documents which satisfy those criteria.

Having selected candidate documents using heuristics, it is possible to estimate and then sum the distance each has to its intended value from the prototype. The sum of squares of each of these will result in the overall error for the final corpus, being analogous to the mean squared error:

$$MSE = 1/n \sum_{i=1}^{n} \sum_{d \in D}^{d} (P_i^d - C_i^d)$$

Where P_i^d is the value of metadata type $d \in D$ of prototype document i, with C being the equivalent selected document that was ultimately added to the results set. n is the number of documents selected.

5.4 Method & Implementation

In order to test the design outlined in Section 5.3, a software tool was produced. The role of this tool is to:

1. 'Read' a seed corpus to construct a description of its metadata;
2. Store this metadata as a description that is transferrable and portable;
3. Replicate the input distribution of metadata by sampling from the web.

This forms the basis of the 'replication', 'rescaling' and 'distribution' scenarios specified in Section 5.2, as well as covering a superset of the features required for some simpler workflows (such as 'repair'). This section details the design and architecture of the solution produced.

5.4.1 Architecture

The tool, named 'ImputeCaT', is a small collection of command-line tools written using the Ruby programming language[28]. It is structured around two main executables: the former of these is called the Corpus ProfileR (hereafter 'cpr'), and is responsible for reading a corpus, extracting its metadata, and serialising the resultant distribution to disk. The latter reads, resamples, and attempts to retrieve documents matching an existing corpus profile—as it is seeking a gold standard, it is named the Golden Retriever (or simply 'gr').

The separation of the tools at that stage offers many advantages: firstly the serialised corpus distribution acts as a file format, allowing a corpus to be passed between users, and secondly it offers a 'DMZ' approach to information security: there is no mechanism for any (potentially sensitive or confidential) information from the original corpus to be used by the retrieval tool, except that which has been explicitly extracted. Finally, this approach permits easy parallel execution of tools using a corpus description for read-only tasks.

Currently, both tools read descriptions of metadata type (and other resources) from a 'profile' file. This provides corpus-design-specific configuration data that would otherwise be impractical to enter through the command-line (for example, mappings from field names to metadata types).

Document Description

Documents are the sampling unit of ImputeCaT, and are used as atoms in sampling algorithms. Each document is represented as a key-value store of metadata 'names' (assigned using the profile description) and their respective values, along with the content that is being represented in the form of a plain text string.

This structure is used even where a document lacks text: building a corpus description from metadata will create a corpus with many documents, yet these will be populated solely by metadata fields and will have empty content elements. Documents are mutable, and as they pass through the system they are liable to be edited in order to refine and change their contents according to a number of processes.

Corpus Description

The corpus description process is structured around gradually adding to a 'corpus' object. This object contains each document represented as a vector of metadata values, $d \in D; m \in d; m \in M$, along with marginal distributions for each (which are stored for optimisation purposes).

The distributions used for each metadata type are arbitrary, and may be controlled in a similar manner to priors in Bayesian inference in order to relax or modify the results. Currently two systems for estimating the empirical distribution of data exist: the simpler of these represents discrete data by simply storing frequencies for each value observed. A continuous form of this represents distribution as a sum of Gaussian kernels—this is designed to disperse the observations and provide a more meaningful overview of the input data. These distributions are

[28] https://www.ruby-lang.org/en/

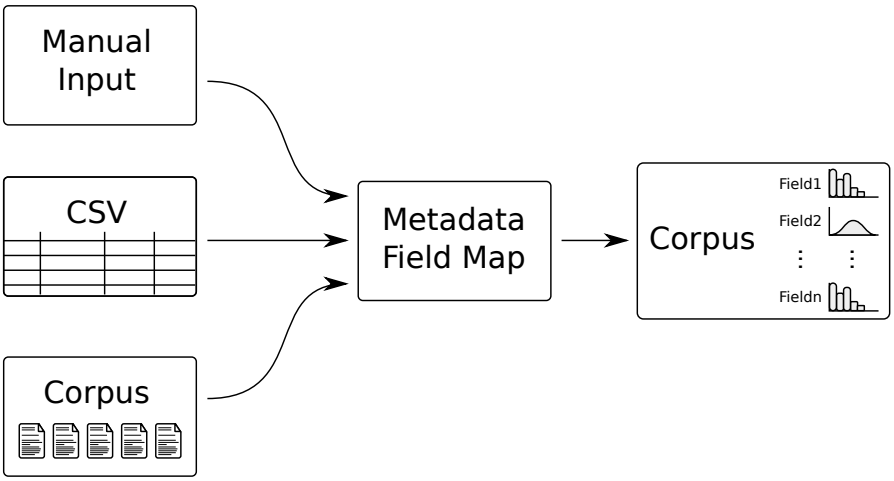


Figure 5.3: The objects involved in forming a corpus description

specified in the profile document provided to 'cpr'.

Each distribution object supports sampling at a given value x, returning $P[x|\text{seed corpus data}]$, and a `random` function that returns a random number following the distribution in question. The discrete distribution implementation uses roulette sampling to implement this, and the smoothed continuous distribution identifies at random one of the contributing Gaussian function then samples from it using the Box-Muller transform[29].

The process of reading a corpus to describe it is one of loading document metadata from files. As illustrated in Figure 5.3, the values of metadata may be loaded from a corpus in CSV format, or from a manual description of the metadata required. In addition to the metadata field values for each document, a mapping from CSV field names to metadata distributions is required, in order to define how each item of data is to be stored and manipulated.

It is possible at this stage to select distributions that artificially manipulate the metadata of a corpus, allowing users to 'filter' or skew an existing corpus without manually editing each file prior to insertion. This could be used to produce stochastic subcorpora, where the probability of undesirable documents is not eliminated but merely reduced, in effect rebalancing the corpus on-the-fly. This is akin to two-phase stratified sampling, yet offers possibilities beyond simple weighting of each input document.

The selection of distributions is performed using the profile file, in a human-readable format. This section of the profile currently maps CSV field names to instances of the distribution objects in question, as shown in Listing 5.1. The parameter passed to the SmoothedGaussianDistribution constructor defines the standard deviation of the Gaussian distributions used, causing slight variation in sampling around each point.

```
1  PROFILE = {
2
3    # ...
4
5    # CSV fields mapped to their distribution type
6    fields: {
```

```
7       'GENRE'          => Impute::DiscreteDistribution.new(),
8       'Word Total'     => Impute::SmoothedGaussianDistribution.new(30),
9       'Aud Level'      => Impute::DiscreteDistribution.new(),
10      'Language'       => Impute::DiscreteDistribution.new(),
11    },
12  }
```

Listing 5.1: Field mappings in the BNC corpus profile.

This approach allows the application of weights and control over the resampling process in a manner that is explicit and more efficient than implementing a different sampling algorithm for each modification.

As features are extracted from the seed corpus, they are assembled into a single document object and stored in the corpus. A corpus is, then, a set of documents with an accompanying set of distribution objects that describe the marginal distributions of each dimension of metadata.

It is possible to design a corpus that has no documents, and is simply a description of desired marginal distributions, however, this approach will limit the power of the rest of the system as it is unable to model any interactions between metadata types. For certain selections of metadata types, this may not be a strong assumption, however, many of the metadata types that can be extracted from existing corpora are designed to be useful to human analysts and are thus non-orthogonal.

It is also possible to import a corpus from text only, by running classifiers and heuristics to extract important metadata during the import. This method requires access to the full text of the input corpus, but has the effect of making any errors symmetric with those of the rejection sampling phase of the retrieval tool.

Resampling

```
1  Selecting random document from #<Corpus:4:9506908>...
2   - dims: 4, docs: 4054
3   - dims: 3, docs: 1652 (Aud Level == med)
4   - dims: 2, docs: 51 (GENRE == W_newsp_brdsht_nat_arts)
5   - dims: 1, docs: 1 (Word Total == 1335)
```

Listing 5.2: Output from the conditional resampler, run on the BNC corpus.

Resampling is the process of producing a document with metadata values that are in some way related to an existing corpus object's distribution of documents.

ImputeCaT implements three different resampling algorithms:

- **Marginal**
  This sampler selects metadata values at random from each marginal distribution, without conditioning upon any values. This will ignore interaction effects between metadata and thus cause overdispersion in the sampled documents, however, it is possible to run this method on a corpus with no documents.

- **Conditional**
  The full conditional sampler selects a metadata dimension at random, selects a value from its distribution, and then produces a sub-corpus conditioned on this value. This process is repeated until no dimensions remain, at which point the metadata values used are returned (as shown in Listing 5.2). Since a sub-corpus must be constructed at each iteration (in order to rebuild the remaining variables' marginal distributions), this process is possible only on corpora with knowledge of individual documents.

- **Partial Conditional**
  This sampler follows the same basic 'selection and conditioning' iteration as the full conditional sampler, yet stops after a set number of iterations. It behaves in a manner similar to that of a blocked Gibbs sampler, producing documents that converge to a similar distribution to the input yet with greater dispersion.

Due to smoothing applied to continuous distributions in a corpus, it is possible to sample at random a value that will return no documents for the conditioning phase of either conditional resampling mechanism: the solution to this is to provide a window width to the method, in which all documents will be selected, effectively providing a reverse function for the smoothing method. Since the standard deviation of each Gaussian kernel making up the smoothed distribution is known, the window size can be set in such a way to guarantee a given probability of selection, for example, a value of 1.645σ would result in a 95% chance of selecting at least one document. Such effects are more likely the more input distributions are artificially processed for editorial reasons. Note that this method is a compromise to make sampling tractable: actual Gaussian kernels extend to $x \pm \infty$, meaning we would need to consider every point in the distribution.

The 'gr' tool uses the full conditional sampler exclusively: this provides the most accurate documents under the assumption that the 'seed' corpus is already fairly complete.

5.4.2 Document Search & Retrieval

After loading the corpus description and profile from disk, the process of iteratively retrieving documents begins. There are a number of different approaches to this task (as detailed in Section 5.3.2): 'gr' uses one that is designed to minimise interactivity during execution, in order to permit construction of arbitrarily large corpora with little effort.

The process of rebuilding a corpus is one of iterating the algorithm below until the size requirements are met:

1. Sample a document from the seed corpus, known as the 'prototype';
2. Using a set of pre-defined search methods, select a set of candidate documents that is a superset of the prototype;
3. Impute missing metadata values using context from the search system and document contents using a series of feature extractors;

4. Identify and select the best candidate document by minimising distance in vector space (rejection sampling).

The approach of scoring each candidate document after retrieval is arguably unnecessary if other retrieval methods are used: for example, manually selecting documents could prove so accurate for some metadata types that it need not be verified automatically (though some annotator agreement measures may result in a similar process to the above).

This rejection sampling method was selected in order to relax the precision requirements for each of the search methods. In practice, each search method is likely to oversample somewhat from a population distribution which is unlikely to fit the seed corpus—the number of candidate documents selected ought to be proportional to the difference between the seed distribution and the population one.

Output is in the form of plaintext, along with stand-off documentation providing details of the metadata for each document. Part of this metadata contains the prototype's details, in order to compute fit offline.

Search Methods

Methods used for searching are implemented as pluggable objects that accept a document to search for (and the names of any metadata field types they respect) and follow some procedure for retrieving candidate documents.

Search methods need not retrieve documents by searching for all metadata fields (though ideal, this is essentially not possible online), however, they should maximise variation in dimensions which they do not deliberately target in order to increase the coverage over the search space. Use of multiple overlapping search strategies in construction of each candidate set ensures that each dimension of metadata is given an equal chance of being selected correctly.

It is desirable that these search methods cover as much of the possible search space as possible, such that they do not apply the bias illustrated in Figure 5.2. Nonetheless, selection of different search methods and heuristics will apply a set of assumptions that must ideally be aligned with the use of the resultant corpus: this is no different to any other corpus construction effort, except that this design explicitly enumerates and codifies this stage. In the ideal case users will be able to select from a library of methods, each based upon empirically-validated theories—application of these would be able to adjust a corpus from one designer's set of assumptions to another, in order to ensure that it best fits a given use.

As documents are selected, their context must be passed on for use by heuristics and other processes. This is accomplished via a special 'meta' entry in the document object, which stores details such as HTTP headers and information on how to continue spidering from the document. Some of the heuristics use these metadata for classification.

A web spidering system, Mechanize [29], is available as a module for use with any retrieval methods. Since it requires seed URLs, it does not constitute a retrieval method itself, however, it is the sole library used to retrieve and annotate web data, leaving behind a handle to the state of

[29] http://wwwsearch.sourceforge.net/mechanize/

the spider in each document—using this it is possible to arbitrarily spider from any document that is currently in the system. This behaviour is vital to retrieval of documents that are correct in only some dimensions, for example, from websites with the correct type and topic but lacking an appropriate word count. Essentially, this mechanism allows other methods to exploit the clustered nature of the web to their advantage in seeking data.

i. Search Through the use of the Azure web services API [30] the Bing search engine can be used to retrieve URLs. These may then be retrieved directly, or used as the starting point for a spider (see 'Hybrid' below).

The Bing search engine allows for multiple search parameters, including language, selection by top level domain, and filtering using a number of heuristics to identify text types (for example, searching for the format of email headers). These techniques allow for lookup of genres, though offer little control over some other features such as document length, which are typically of less interest to the average web user. Nonetheless, for metadata which is searchable, the method of seeking a prototype offers a way to cut through the distortion applied by the search algorithms[31].

The search mechanism in 'gr' operates by detecting languages from metadata, searching the appropriate 'market', and uses a series of pre-computed keyword lists to find matching genres.

Keywords for the system are passed as an argument in the corpus profile, and can thus be changed per-design to fit the classification used. Both of the corpora used for testing in this thesis use Lee's BNC classification, for which log-likelihood-based keywords were computed. Keywords are identified for use as search terms weighted by log likelihood values.

ii. Directory Search By using an existing web index, it is possible to map taxonomies directly to sets of links. This has the distinct advantage of reliability and replicability, yet significantly reduces the potential pool of documents for selection.

The ability to use different indices (or search them) based upon another dimension of metadata makes this method suitable for use with very special-purpose corpora. For example, one may search Project Gutenberg (a large repository of free ebooks) where the text type is 'book', yet return to Bing search or DMOZ [32] for less specialist documents.

Some of these resources may also fall offline, for example, a complete dump of Wikipedia and DMOZ are available and are used directly by ImputeCaT to find URLs.

iii. Hybrid The hybrid approach uses a search term or directory entry as a set of seed URLs from which to spider.

The most naïve implementation, implemented here, simply spiders a certain depth from each search URL. It is also possible to select URLs to follow based on the prototype's values, either to seek a document directly or to maximise variation across non-controlled-for dimensions.

In his examination of different sources of URLs, Barbaresi identifies a lack of diversity in search engine results relative to sources such as DMOZ[11], whilst also noting that such directories

[30] https://azure.microsoft.com/en-us/
[31] Though this method is still subject to bias in their spidering algorithms.
[32] A.K.A. The open directory project, available at http://www.dmoz.org/

underperform when applied to certain languages. The prototype-seeking behaviour detailed here has the benefit of being able to switch between such approaches depending on language, maximising coverage for a given URL source.

iiv. Existing Fringe This mechanism simply searches the existing set of documents, left over from previous searches, for the closest match. As all candidate documents are pre-ranked, this method requires no action. It is equivalent to policies in genetic algorithm selection that retain portions of the population between generations.

Each relevant document in the fringe may also be used as the starting point for a spider, relying on the clustered nature of the web to maximise precision of retrieval.

Post-processing and Context

After download, candidate documents must be converted to plaintext. Processing modules are chainable, meaning that it is possible for each search method to construct a standard pipeline to normalise any contextual information and perform tasks such as boilerplate removal prior to document ranking and imputation using the heuristics.

i. Boilerplate Removal Boilerplate removal is performed by JusText[139]. Though currently this uses the bundled English stoplist, it would be possible to use the document context language attribute to automatically select a stoplist depending upon the input document.

Heuristics

Hailing from multiple sources, documents are unlikely to be annotated with sufficient metadata to permit comparison to the prototype. It is thus often necessary to impute the values of metadata from what external variables are available (for example, meta-tags or HTTP headers) and the document content itself.

Unlike search strategies, heuristics need not consider the practical concerns of data sources, and are thus easier, simpler algorithms, often comprising simple classifiers or counts. Nonetheless, error in these routines sums with other sources of dispersion: mechanisms for measuring the overall error of corpus documents (detailed below in Section 5.4.2) conflate these two sources.

Heuristics are also used to determine the 'most similar' document to the prototype: each heuristic object harbours two methods: one to impute the value of a metadata key based on the document contents, and one to measure distance from one value to another. For many heuristics this is a simple normalised difference between two numbers.

i. Genre Genres are imputed through the use of a keyword-based classifier. This is one of the more challenging dimensions to impute due to its complexity, the importance of accurate detection (and thus corpus replication), and the ambiguity of classifications in many genre taxonomies.

The representation of genre used within ImputeCaT was selected to align with human-usable taxonomies, and as such is largely aligned with the BNC index[110]. Selecting this taxonomy burdens any execution of the tool with various assumptions about the data and the desired output that therefore align with Lee's genre distinctions (and any inadequacies thereof). This is a perfect example of the value of explicitly selecting and stating one's design through the corpus profiling tool.

The genre heuristic in this implementation uses keyword data to classify texts based on their content. This is a rudimentary but reliable approach that doesn't require any consistency in web data retrieved (but also thus ignores any useful meta-information therein).

Rather than simply assuming the taxonomy used elsewhere, it would be possible to represent genre arbitrarily, either as a set of entropy measures against certain features, or in a manner more fitting of the source data. This approach is valuable where the source documents are less well defined, or where particular control is needed over the composition of a corpus (such as frequency profiles for certain words).

One particular use of this approach would be the modelling of under-resourced languages using bottom-up detection of features, as in Sharoff's 2007 study[160]. As with all such approaches, however, the reliance on automatic feature detection strips away a large portion of the utility of genre definitions, in that they are often less easily understood by humans.

There is some argument for this approach, however, when compared to the often context-based labels applied to the BNC. The classifier in ImputeCaT is tasked with accurate identification of classes that are not particularly easy to separate by linguistic features. One possible reason for this is their tendency to conflate topic with form: `W_email` and `W_hansard` are largely stylistic differentiations compared to `W_religion` and `W_ac_medicine`. The resultant classifier, following a multinomial naïve Bayes approach[124], is described in more detail in Chapter 6.

The difficulty with which genre is classified makes a case for careful selection of metadata, and illustrates a compromise: often, the most operationalisable and useful metadata selection will not be easily understood by computers, and yet more synthetic features extracted from texts are often difficult to seek in an online environment that is designed specifically for human use.

Flesch reability score	BNC Audience level	Standard Deviation
63	Low	14.4
55	Med	12.7
47	High	12.4
82	(unclassified, speech)	20.6

Table 5.1: Target Flesch reading ease scores and their equivalent BNC categorisation.

ii. Audience Level Reading ease metrics are used to estimate the 'Audience Level' classification provided by Lee. This takes the form of the Flesch reading ease score[51], quantised through selection of the nearest target value as in Table 5.1.

Distance is normalised on this score by dividing the distance between classifications by the maximum distance possible between all classifications. The score is thus output in terms of

distance between categories: comparison of any 'low' score to any 'high' one will be equal regardless of the degree to which the original value is low or high.

Note that this categorisation is a design choice on behalf of the corpus builder: it is also possible to code the reading ease measure as a continuous value, which would offer more nuanced measurement of similarity.

iii. Word Count Words are counted by simply splitting the plaintext string on whitespace, and counting the parts.

Whilst this imputation algorithm is trivial, normalising the distance metric is not: the normalised distance between words score is provided by artificially imposing an upper bound beyond which the distance is simply registered as 1.

Measurement of Error

Notably, by scoring each document according to its target, it is possible to measure overall error for each metadata property as the corpus is being built. In a unidimensional context this greatly simplifies the process of minimising error (since it is possible to adjust the seed distribution to compensate for errors on-the-fly), however, such comparisons would be intractable when applied to the intricate conditional distributions found in real-world corpora.

As documents are selected, it is possible to compute their sampling error by taking the difference between metadata values relative to the prototype. This difference will be in units defined per metadata type, following a distribution that, in the ideal case, is uniform.

The error surface is, in reality, as intricate as the original distribution of documents: there is no way to compute errors for all conditions within the corpus without simply comparing documents. To operationalise measures of total distance, therefore, we may summarise the error in a number of ways:

- **Mean Error**
  This is simply the mean of all errors for each metadata item, and provides a coarse-grained measurement of the overall bias for a given dimension. This measure is analogous to measuring the area under the difference between marginal distributions of the input and output distributions for a given metadata dimension.

- **Mean Squared Error**
  A sum of the squares of differences between each document and its prototype, this provides a dispersion measure that does not include direction. This measure is commonly used to assess the quality of sampling, however, may not adequately represent the difficulties of sampling data online.

- **Uniformality of Residuals**
  Correlation between prototype and document metadata values evidences the responsiveness of search methods, and is testable using Mann-Whitney's U.

If weighting the candidate document selection function, greater mean standard error should be observed on those dimensions with lower weights.

5.5 Summary

This chapter has detailed the design and implementation of a method that aims to extract and summarise corpus data in a human readable format. This data can be modified in a number of ways before forming the seed of a new corpus, which uses data sampled from arbitrary sources to construct a clone with known bounds for error.

This approach is beneficial compared to more conventional bootstrapping as it permits inspection and adjustment of metadata at a level relevant to the use a corpus is likely to be put to: the external metadata. This also allows us to inspect and manage bias somewhat, under the assumption that we are able to translate the meaning of metadata in the context of the seed corpus to another source of documents (such as a search engine).

This mechanism is a workaround for Moravec's paradox[132]: a compromise between the low-level, unambiguous features needed to represent a corpus accurately to a computer and the high-level, biased mechanisms of selection online, which are oriented towards providing a commercial service to humans. Careful selection of said metadata is thus crucial to ensure that this balance is respected in the resultant corpus, particularly with respect to any aspects that should be studied. It is the intention of the design detailed here to make that process explicit, replicable, and repeatable: where this is not possible, the sources of error are documented.

The implementation detailed herein constitutes a single workflow, the dissemination and replication of a corpus. This workflow was selected as it is largely a superset of the others, demanding accurate description and automated identification, retrieval, and assessment of potential documents in a 'turnkey' manner.

The use of searching and keyword lists in document retrieval bears comparison to BootCaT and other 'seed' based approaches—indeed, the method here presents a generalisation of that used in BootCaT. If using keyword search for retrieval, it is possible to emulate the behaviour of BootCaT by ensuring that:

- The input categories (for which keyword lists exist) are the source of the keywords used;
- Heuristics must impute their category selection based only upon the same raw data as the above;
- Only the search-based data sources are used;
- There is only one dimension, i.e. the one which is reliant on keywords for retrieval.

In addition to these, the framing of an input corpus in terms of human-usable metadata means that a measure of retrieval error is possible for ImputeCaT, something that must be performed from the resultant keywords in those systems working bottom-up[33].

[33] Though ultimately these constitute the same process, there is value in translating it through a human-readable lens.

Chapter 6

Evaluation & Discussion

The evaluation of corpus construction is a particularly tricky area: without a gold standard or real-world task to frame the results, it is often difficult to tell quantitatively which differences between corpora constitute improvements.

The way seed-based methods work offers an ideal source of gold standard data, nonetheless, only experience and repeated use can reveal the manner of the differences identified, and its impact on experimental results. This evaluation will use this reflexive method to identify overall responsiveness to the seed corpus, along with an examination of individual components of the system as a method for revealing specific strengths and weaknesses.

This evaluation is based around one of the tasks identified in Chapter 5—that of rebuilding a corpus from scratch based upon a seed's proportions. This task is effectively a super-task of reconstructing or repairing a corpus, and is equivalent to many scenarios involving disseminating sensitive corpora with minor restrictions on the metadata dimensions used.

This chapter begins with a description of the use cases covered, and a detailed rationale of how these form testable, objective research questions.

Section 6.1 details the two main approaches to evaluation: those applied to individual components, and those applied to the results of running the implementation as a whole.

Following that, each stage of the corpus building process is evaluated in a white-box fashion, starting with the heuristics used to classify documents (Section 6.2) and moving on to the process of identifying 'target' prototype documents in vector space (Section 6.3), and then retrieving documents from the web itself (Section 6.4).

Finally, the implications of each component's functionality to the final process, and what the results tell us about the state of the web as a document source for resampling BNC-like corpora, are discussed in Section 6.5.

6.1 Evaluation Criteria

The approach used in this evaluation is one of treating the system to be tested as a method of replicating the features of a corpus, passing statistical properties of the input corpus through to

the output.

The success of this method is dependent on the structure of the input corpus, and the structure of the population new documents are drawn from, and as such this evaluation is just a single datapoint in the space of possible research questions. If testing with another seed corpus, one must change the search strategy modules to fit, changing the overall results of the tool. Essentially, this is a manifestation of 'garbage in, garbage out': moving search strategies into the implementation merely makes this more explicit. For that reason, this evaluation uses an extremely popular corpus, the BNC, to ensure some degree of transferrability.

One method of testing this is to see the output corpus as a model of the input, given certain assumptions that include the selection of metadata types and search strategies: providing those assumptions hold, much of the variation in the input corpus ought to be explained by variation in the output. Measuring differences between the two allows us to identify specific areas of poor fit, which can then be improved either by altering the pluggable modules to fit the ground truth.

The purpose of this evaluation is to test both the search modules used for the particular corpus given, and the overall method: if no set of reasonable assumptions can be found, this is an indication either that the population of documents online is fundamentally different to that of the input corpora tested, or the method presented here is incapable *by design* of identifying appropriate documents. Differentiating between these two is not possible quantitatively.

This evaluation will centre around replication of a large general-purpose input corpus that is sampled at the document level (that is, each datapoint is a document rather than a single word or sentence). This design has been chosen as it represents the most common current design and because metadata at the document level are consistently both meaningful to humans and well researched. It is also similar to the approach taken by existing search methods for building corpora, making results indicative of search engine behaviour.

Research questions surrounding this method run to:

- **Components**
  How valid are the assumptions of each of the retrieval methods and heuristics selected?

- **Overall Application**
  For general-purpose input corpora, to what extent (and in what manner) does the output corpus resemble the input 'seed' corpus?

- **Residual Variance**
  What consistent features remain variable between the output and input corpora, i.e. what data cannot be sought online using the heuristics/search methods selected?

The error of each heuristic component is contributory to the excess dispersion in the output corpus. By design these modules are unambitious, relying on existing methods and tools already tested in the literature, however, their performance upon the data used here will be evaluated in order to better explain sources of error. As this system remains a proof of concept, the selection of these modules is limited.

6.2 Performance of Heuristics

The heuristics selected for this evaluation are formed around Lee's BNC index[111]. This selection was chosen because of their alignment to operationalisable, human-level metadata and the existence of multiple corpora with this level of annotation.

There are two main approaches to populating the corpus description using these heuristics: either read the seed corpus' contents and classify each data point, or read a list of metadata from an existing index. The latter approach is used in this evaluation, since it is applicable to corpora with partially-missing data (such as the personal corpus data resulting from data gathering in Chapter 4).

The accuracy of the classifiers listed here is responsible for minimising excess dispersion relative to the input corpus. The nature of their residual error is also going to apply bias to the resulting data set.

Since many of these heuristics surround operationalising a corpus, a large body of research exists for classifying and extracting useful dimensions from texts. The heuristics presented here are proof-of-concept only, and it is expected that the design of the heuristics used for a study is selected to match the theoretical basis of any analysis.

In most corpus designs, word count would be considered a measure of the size of the corpus (rather than a property of its constituents). The method evaluated here is capable of retrieving IID samples at different levels, and demands a different selection of heuristics and metadata when operating at the word or sentence level. Document level metadata are both high-level enough to be distributed for confidential corpora and descriptive enough to enable accurate retrieval (by contrast, word or part-of-speech frequencies would reveal much of the contents of the original corpus, which may not be desirable). They are also aligned with storage and indexing, which are both often performed at the document level online.

6.2.1 Audience Level

The BNC User's Reference Guide [30, p. 14] describes audience level thus:

> *. . . a subjective assessment of the text's technicality or difficulty.*

This is expanded upon slightly by Lee [110, p. 68], saying:

> *Audience level, on the other hand, is an **estimate** (by the compilers) of the **level of difficulty** of the text, or the amount of background knowledge of its subject matter which is assumed.*

One approach to modelling this complexity is to use word lists such as those used in education, however, this is difficult to apply to such disparate topics without extensive compilation of such lists. A simpler approach uses metrics computed from the morphology of words to form a readability score. Such metrics are already widely used in word processing software and for designing documents for public consumption.

The classifier used for evaluation is based on the Flesch reading ease score[51]. This score is widely used, easy to compute, and correlates with the simple 'low/med/high' classification used in the BNC metadata:

$$readability_F = 06.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

In order to establish the meaning of the BNC audience level categories in terms of this score, means were computed from the BNC world data. These means, shown in Table 6.1 (a reproduction of Table 5.1), were used by a simple classification algorithm that selects the category with the nearest mean value. This naïve method yields the accuracy indicated in the 'naïve acc.' column of Table 6.1, indicating that the readability score is not an ideal measure of the audience level.

$readability_F$	BNC Audience level	Standard Deviation	naïve acc.
63	Low	14.4	65.9%
55	Med	12.7	25.0%
47	High	12.4	66.3%
82	(unclassified, speech)	20.6	-

Table 6.1: Target Flesch reading ease scores and their equivalent BNC categorisation.

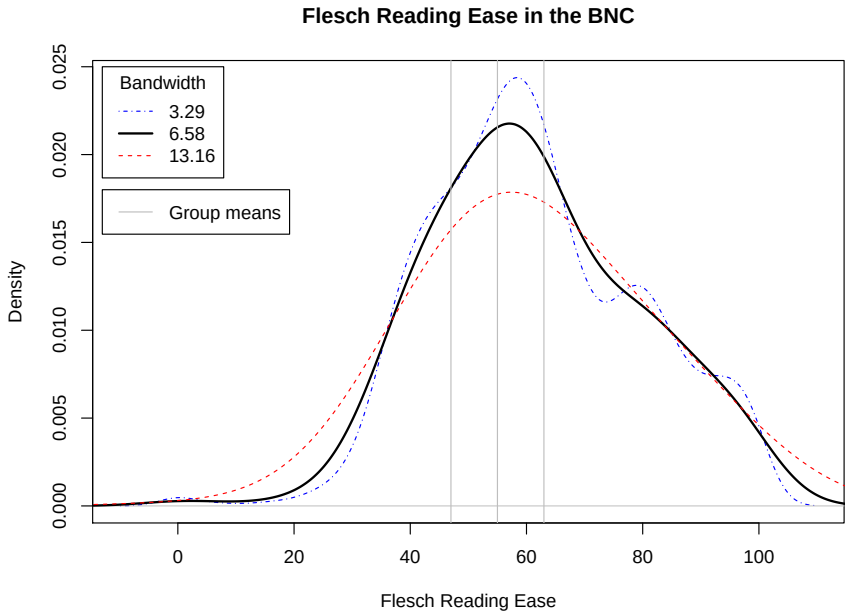


Figure 6.1: Flesch reading ease distribution in the BNC.

Rather than continue to use the categorisation used in the BNC, each text was scored for readability and the raw Flesch reading ease score used as its point within that dimension. The 'cpr' profiling tool then simply represents the underlying distribution as a continuous one (rather than artificially discretised into 'high', 'medium' and 'low').

The standard deviation of categories within Table 6.1 provides us with a convenient source for the bandwidth of the smoothing function used to identify the empirical distribution of

the reading ease scores. Using graphical methods (illustrated in Figure 6.1) the value of $\sigma/2$ was chosen to represent the distribution: this offers a tradeoff between accurately portraying the distribution and permissively selecting documents online. This method also clarifies the difficulties associated with attempting to accurately classify audience level based on readability metrics.

Where a text is unavailable, knowledge of the standard deviation of each of the readability categories in the BNC allows us to produce an approximation of the distribution of readability scores. This approach (as well as the smoothing above) illustrates the value of manually modifying the text distributions during the design phase.

As we are using a proxy variable to impute the reading ease, any classifier accuracy issues are deferred to the assumption that 'reading ease' is something we, as corpus designers, care about.

6.2.2 Word Count

Word counts are, for our purposes, trivial to compute. There are two challenges to operationalising word counts, however.

The former of these pertains to metadata and other boilerplate within files. Files must be processed prior to word counts being performed. This means (like many other tools) that we must post-process candidate files even if we elect to discard their contents.

Secondly, it is necessary to normalise deviance measures when comparison documents: this poses a particular challenge as word counts are open-ended. The method used here was to enforce an arbitrary limit, above which any distances in the word count dimension register as 1. Because the sampling method does not primarily rely on gradient descent, this is fairly safe, however, selection of the threshold does impact the importance assigned to word count (vs. some other metadata dimension): too high and it will undervalue differences in word count.

The arbitrarly limit selected for the BNC was $10,000$. This is sufficiently large that any difference greater than it may be considered 'absolute', that is, if there is a difference in length of over $10,000$ words, we can reasonably conclude that the lengths differ enough to be describing different texts. In practice (as we shall discuss in Section 6.4) most documents fall short, and this limit could be reduced.

It is worth noting the difference in philosophy this method embodies compared to many corpus studies: the word count here is a property of each document, and *not* a measure of corpus size. Unlike many methods, we are sampling at the same level we are presuming analysis at.

6.2.3 Genre

Genre is arguably the most important single stratification applied to a general-purpose corpus (even forming the bulk of the definition of its name), yet it remains one of the more nebulous terms. Here we follow Lee's definitions once more, commensurate with the BNC index that serves as the source of much of the metadata for our corpus definition file.

Due to the complexity of genre taxonomies, and thus the relative difficulty of classifying documents therein, there is significant motive to take the same approach as audience level above:

using a self-defined or simpler taxonomy to simplify imputation or selection. On the other hand, the BNC classification is well-known and likely to be used in any corpus augmentation task (e.g. "I want this subset of the BNC, but *more* of it."). We describe here an approach using the existing BNC classifications for this reason, with a discussion of the implications of other options below.

As this is a proof-of-concept system, the accuracy of the classifiers used need not be cutting-edge. This relaxed requirement is reinforced by the notion that any errors are 'sane', falling along such ambiguities as identified by Lee [111, p. 11]:

> *It may be the case that the actual content/topic of these linguistics-related texts makes them seem less like social science texts than arts or applied science texts. . .*

This 'sanity of error' is enforced by using a distance measure that is based on rank correlation of token type frequencies, allowing two categories to be judged on a scale more meaningful than simple boolean equality. This luxury may not be available for all potential discrete classification systems (such as those read from external indices), and is not a requirement of convergeance to the input distribution.

The distinction between genres is often somewhat variable: some, such as `W_newsp_brdsht_nat_*`, are made between the subject of texts, whereas others fall along lines of context or what others may term 'text type' (`W_email` vs. `W_essay_sch`). This makes it particularly difficult to identify a model that linearises features within the dataset consistently enough for many classification algorithms to work well.

Lee's taxonomy is partially hierarchical, with many of the more detailed categories harbouring both this 'text type' distinction as well as one regarding topic. Both were retained here as they align to two distinct processes when sampling: the text type is an indication of (roughly) where a document may be found, and the topic regards more which document to select.

Classification The problem of identifying BNC genre from free text is a simple 70-way classification task[34]. In practice, the subjective nature of the taxonomy and the large number of classes makes this a challenging endeavour. This is a clear illustration of the difficulties encountered using the ImputeCaT method: whilst it may not be necessary to find documents online with great accuracy, there must exist *some* method of discriminating between them in a useful way.

In his 2007 paper on web genres, Sharoff[160] fits a number of classifiers to the BNC, using a variety of genre taxonomies. Therein he reports success using a supervised learning approach to classify 'grouped' versions of the BNC taxonomy: one aligned to the EAGLES[166] text typology recommendations, and one 10-way grouping based on unsupervised methods.

An alternative method would be to use entirely ground-up techniques such as factor analysis or LDA. These mechanisms will extract maximally homogeneous groupings from text, but at the expense of predictable alignment with human understanding (and thus any research questions or search engine designs).

Since part of the intent here is to retain a parsimonious corpus profile document, this heuristic makes use of the original categories from the BNC taxonomy. Nonetheless, Sharoff's study was

[34] Such that a 70-way classification task may be described as simple at all.

used as a starting point for the methodology, in line with the goal of using established methods for heuristics.

The ambiguity and external nature of some of the distinctions drawn in the BNC index causes problems here, and represents a common dichotomy in linguistics: a human-readable and intuitive classification is rarely easy to model quantitatively (to the point where we are often simply unable).

Ultimately, the classifiers detailed here used the naïve Bayes approach and were fit using WEKA[69]. In accordance with the findings of McCallum & Nigam, the multinomial model was used over the Bernoulli due to its ability to maintain classification accuracy in models with large numbers of dimensions[124]. This also unexpectedly yielded the benefit that it was much faster to fit.

For the use-case presented here, absolute accuracy of the classifier matters little: we care only that 'positive'ly classified cases are reliable for any one pass over the set of candidate documents, *not* that every candidate is selected. Because of this, and because the number of classes is high enough to indicate that errors are unlikely to occur in the direction of the current class, the evaluations below focus on the precision of each method.

As ever, the classifier performance reveals as much about the taxonomy of the input data as it does the classification method. Presented here are six models that cover different subsets of the BNC:

- **BNC70**
  A classifier built using all 70 categories of both written and spoken BNC portions.

- **BNC69a**
  This classifier omits `S_conv`, which was deemed unreasonably general in its content and thus difficult to retrieve.

- **BNC69b**
  The BNC70 classifier, omitting `W_misc` due to its breadth.

- **BNC68**
  Omitting Both `W_misc` and `S_conv`.

- **BNC46**
  Written portions of the BNC, containing `W_misc`.

- **BNC45**
  Written portions of the BNC, without `W_misc`.

The decision to omit sections of the BNC was deemed a necessary trade-off to achieve functional performance: for the purposes of this evaluation, the actual subset used is only of importance to the generality of the results (rather than their veracity). Likewise, the priority is not to produce a genre classifier, but merely to use one within the larger method: simply, it is more important that the classifier is functional than that it is general.

Table 6.2 shows the resultant accuracy figures for the models trained on the BNC subsets mentioned. 10-way crossvalidation was used to estimate the above statistics, in an effort to

Corpus	Accuracy	Precision	AUC
BNC45	70.7%	0.721	0.901
BNC46	63.7%	0.666	0.877
BNC68	71.5%	0.715	0.913
BNC69a	65.9%	0.685	0.893
BNC69b	72.1%	0.732	0.918
BNC70	66.6%	0.689	0.899

Table 6.2: Weighted mean performance (per class) for each classifier.

reduce the potential for overfitting to a fixed test/training set. The statistics are formed from a weighted mean of each classes performance against all others.

These figures reveal that *s*poken conversation (`S_conv`) has the most deleterious effect on classifier accuracy, indicating that it overlaps most with other categories. Whilst it was expected that `W_misc` would also do this, the performance improvement shown in BNC69b over BNC68 indicates that `W_misc` contains some distinct features that are best retained. Though using a different design of classifier, overall performance is roughly in-line with Sharoff's BNC-based classifiers[35][160]. The 45- and 46-class models are written-only equivalents, and are listed here due to their use in the retrieval task below.

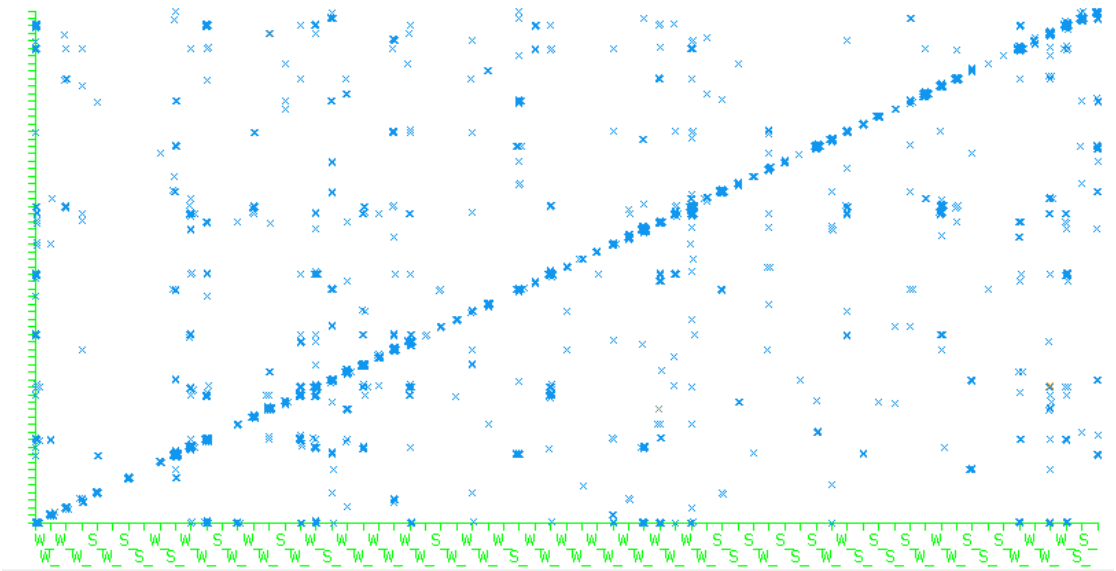


Figure 6.2: BNC68b confusion matrix

There are also potential arguments for the omission of many other categories from the taxonomy used here, however. Figure 6.2 shows the confusion matrix for the BNC69b classifier—note that some jitter has been applied to the plot to accentuate the overplotted items. This shows a rough smattering of difficult-to-classify classes, though a more detailed inspection of the per-class tests on the model indicates that only two of these perform with AUC values below 0.7: `W_fict_drama` and `W_essay_univ`, which are classified in a way indistinguishable from chance. Some classes with distinct lexical features (such as `W_email` and `S_sermon`) are classified particularly accurately.

[35] Attempts to fit SVM-based models using RBF kernels lead to models with accuracies around 51% (AUC ≈ 0.74) for the 68-class model.

The error surface described in Figure 6.2 will contribute to the resulting error of the system, and as such further increase the variance of the output corpus. Again, this is analogous to the coding stage of any conventional corpus building project, but with the distinct advantage that the propensity for classification error is well-known. As shown, there are no instances where a genre is consistently mis-classified, indicating that errors will not systematically bias any output to a great degree. The full accuracy data is presented in Appendix D.

One final note should cover the generality of the methods used here. As with all portions of the heuristics, the choice of classifier is heavily theory-laden, and should conform to the underlying data's nature as much as possible. Classification of genre is a particularly awkward aspect since it relies on imputing fairly complex external information (such that it can then be worked with by the software) from linguistic content, something that, as a rule, should be avoided for the sake of statistical validity. Notably, the EAGLES guidelines on text typology[166] warn:

> *The classification of texts into different genres seems to have been mostly achieved through external criteria rather than through internal, linguistic criteria.*

The manner of retrieval and the nature of the web offer some avenues for working with external data and closing this gap: It is firstly possible to retrieve documents according to an existing genre directory, all but guaranteeing a fit. Further, HTML and HTTP metadata and stylistic properties of web documents (or those which are closely linked) may indicate genre as effectively (or more) as linguistic content. These are all places where further research dovetails with the methods described here to increase overall accuracy. Further work would be necessary to validate methods of reliably extracting and using this data, especially where it is patchy or difficult to extract. Such efforts would prove particularly fruitful, however, where the seed corpus originates online, allowing the profiling and heuristic classification stages to operate using the same implementation.

6.3 Performance of Resampling

The accuracy of the resampling process depends lagely on two user-controlled properties:

- The amount of smoothing applied to continuous metadata in the corpus. Since values are selected from the smoothed corpus values, it is possible to select values that are non-identical to the input corpus. This is generally a desirable property, and the kernel function used to smooth the input data is user-defined and has known statistical properties.
- The number of documents selected. As this increases, the overall distribution of the output corpus converges to that of the corpus description file.

The intricacy of the input distribution largely defines the necessary size of the input corpus: an input corpus that is defined only in terms of simple marginal distributions is simpler to

reproduce, that is it contains *less information*, than one built using the full conditional distribution of a large, general corpus such as the BNC.

The question of when to stop sampling documents is related to the problem of corpus size in general: the output corpus is conceptually a model of the input corpus, and should contain enough data to be representative of the relationships therein whilst following the same guide. This is best assessed by measuring the uniformity of residuals. A suitable end condition for many uses of the output would thus be a combination of corpus size and residual uniformity (notably it is possible to constantly balance the uniformity of residuals during the rejection sampling phase too).

The evaluation here establishes the resampling algorithms' ability to produce copies of the input corpus with uniform residuals, and establishes a 'best case' baseline against which the results of document retrieval may be compared.

The research questions answered by the analysis here run to:

1. Does the resampler converge on the same distribution as its input?
2. How many 'target' documents are required to converge upon the input distribution (at some known probability)?

6.3.1 Method

Evaluation of the selection method is possible separately to document retrieval as the input distribution is known. This means that, unlike a bootstrapping operation, we can rely on model deviance measures (rather than heuristics such as autocorrelation) to indicate the point at which we have sufficient data to represent the input distribution. Since documents are retrieved and compared against their 'target' as selected by this process, it is thus possible to guarantee conformance to some property of the input distribution, providing retrieval is performed accurately. The eventual error for the corpus is a sum of these two deviances.

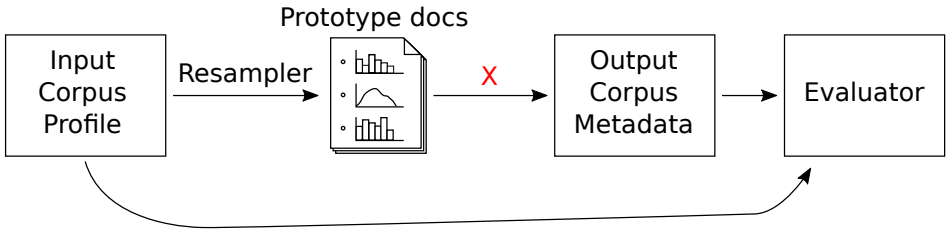


Figure 6.3: Data flow for evaluation of the sampler.

The data flow outlined in Figure 6.3 is identical to that used for the final selection, with the omission of a retrieval stage at x. This means the evaluation will be performed under the assumption that the retrieval process is always able to find a suitable document. The operation performed by the evaluator is essentially a comparison between two distributions, and may be performed using any number of algorithms with the one requirement being that it can practically

be executed after each document is retrieved (note that for the purposes of this evaluation, such a requirement is less critical).

Evaluation methods provided in the prototype implementation are:

- **Mean Error**
  For each dimension, the sum of the deviance from each document to its target value is divided by the number of documents in the collection. This provides an asymmetric form of the commonly-used Mean Squared Error.

- **Distribution Comparison**
  A commonly used method for comparing empirical distributions, this is computed by taking the differences between the cumulative density of each dimension's data and is thus less sensitive to order than MSE methods. Typically these methods are able to provide confidence bounds related to statistical significance.

- **Linear Modelling**
  A linear model is constructed and fit with parameters according to the input dimensions. If this model proves statistically equivalent to the null model, then the residuals are considered uniform.

For the purposes of this evaluation, a distribution comparison method, Log-likelihood, was selected due to its known statistical bounds and wide use in corpus linguistics[36]. The major disadvantage of using this method, that it is harder to directly relate to the content of each text, is not relevant at this stage of analysis.

Log likelihood comparison methods are already widely used in corpus linguistics for keywording and corpus comparison, and measure significance (using Wilks' theorem[183]), rather than effect size (as with many information-theory-based deviance measures such as MI and Jaccard). This evaluation makes use of such properties in that we desire to know at what point the output distribution is, to a given standard of probability, drawn from the same population as the input.

This comparison of the output and input distributions was repeated as documents are re-sampled, the point at which the two cease to be significantly different noted. This process should establish both RQs:

1. Load input corpus;
2. Repeatedly sample from the input corpus, inserting documents into the output corpus;
3. After each document, compare the output distribution to the input.

Note that this method treats each metadata dimension of the input corpus as an independent distribution, effectively evaluating the joint distribution of the sample. This was done to permit comparison where obtaining a sample large enough to provide sufficient statistical power is largely impossible—even in the BNC, there are seldom many documents that share *all*

[36] It it worth noting that the log-likelihood statistic is just one of a large number of valid goodness-of-fit tests, the main contender being two-sample Kolmogorov-Smirnov

metadata values. This has the effect of weakening the test such that the 'Conditional' sampler is indistinguishable from the 'Marginal' sampler (as mentioned in Section 5.4.1).

6.3.2 Results

The evaluation here focuses on categorical values with over five members due to the asymptotic nature of the log-likelihood statistic making it unreliable below this number. As such, the word count was grouped into 17 bins, and the true BNC readability category was used rather than the (continuous) Fleisch reading ease score. This results in three variables that must converge prior to similarity being determined:

- **Genre**
  The genre category, as fit using the classifier heuristic mentioned elsewhere, only using all 70 categories;
- **Words**
  The word count, binned to form bins with no fewer than five texts in each;
- **AudLvl**
  Audience level, categorised as 'high', 'med', and 'low';
- **Mode**
  Spoken or **W**ritten.

The word count is binned irregularly due to the low frequency of files with more than 80,000 words listed in the index. This results in a distribution shown by the histogram in Figure 6.4, with the minimum bin frequency being 8. Other variables were left unmodified.

The convergence rate of each of the variables is, as mentioned above, dependent on the information content of the input bins' frequencies (due to the complexity of the sample) and the number of bins with fewer than five members (due to log-likelihood converging slowly to the χ^2 distribution). Each variable's convergence may thus be appropriately expressed in terms of how long the average bootstrapping process takes to converge upon the distribution to a given confidence value, relative to the size of the input.

A basic description of the bins used for each input metadata item is presented in Table 6.3. Medians and interquartile ranges were used as centrality and deviance measures due to the Zipfian nature of the bin distributions (a feature expected to manifest in most corpora). The 'ID' column gives the 'index of dispersion'—a normalised measure of the variability in the distribution's bins computed by dividing the centrality measure (median) by the dispersion measure (Inter-Quartile Range).

Metadata	bins	$\lvert n_{bin} < 10 \rvert$	$median(n_{bin})$	$IQR(n_{bin})$	ID
Genre	71	16 (22.5%)	26	49	1.88
Words	18	2 (11.1%)	117.5	359.25	3.05
AudLvl	4	0 (0%)	869	317.5	0.36
Mode	2	0 (0%)	2027	1117	0.55

Table 6.3: Distributions of metadata dimensions in the input corpus (BNC index)

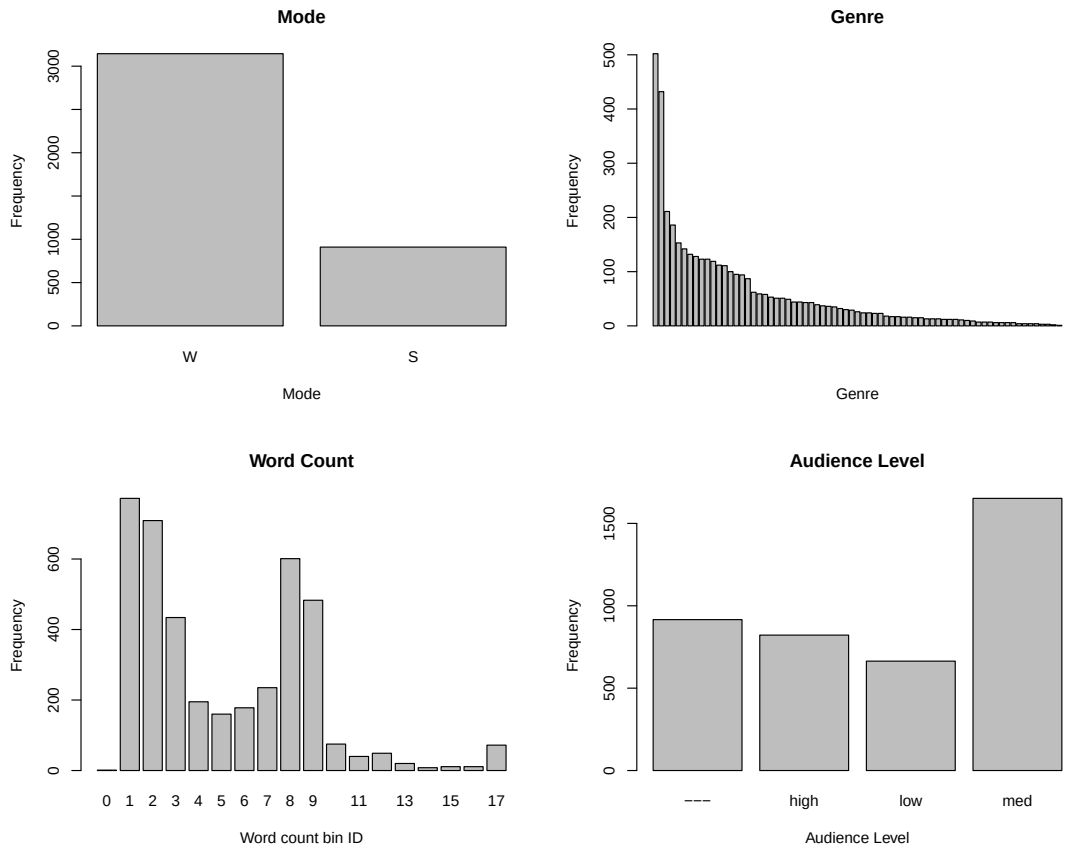


Figure 6.4: Word count distribution from the BNC index (n=4,054, breaks at 5k < 80k, max=428,248)

The figures in Table 6.3 represent common cases for real-world corpora:

- 'Mode' is a simple binary classification with very different group sizes;
- 'AudLvl' is a simple classification with a relatively 'flat' distribution;
- 'Words' is a bimodally distributed measure with a large amount of variation but also very large numbers in each bin. This is a good representation of the counts of many linguistic features;
- 'Genre' is moderately heterogeneous, yet has a distribution with many low-frequency items in it.

We may reason from the index of dispersion that convergence will occur faster for 'AudLvl' and 'Mode' than for 'Genre' and 'Words' — the ordering of the other two largely being determined by the conservative estimation of log-likelihood for low-frequency bins.

Figure 6.5 displays the log-likelihood score of the output distribution as it is built. These graphs clearly illustrate the relative complexity of the input distributions, as well as the downsides of estimating log-likelihood measures for low-frequency results. After an initial burn-in period, all dimensions gradually converge upon the target distribution. For any input distribution with low counts, the similarity measure will not be accurate until these lowest bins are filled—marked on the graph roughly as the point where the output distribution is equal in size to the input.

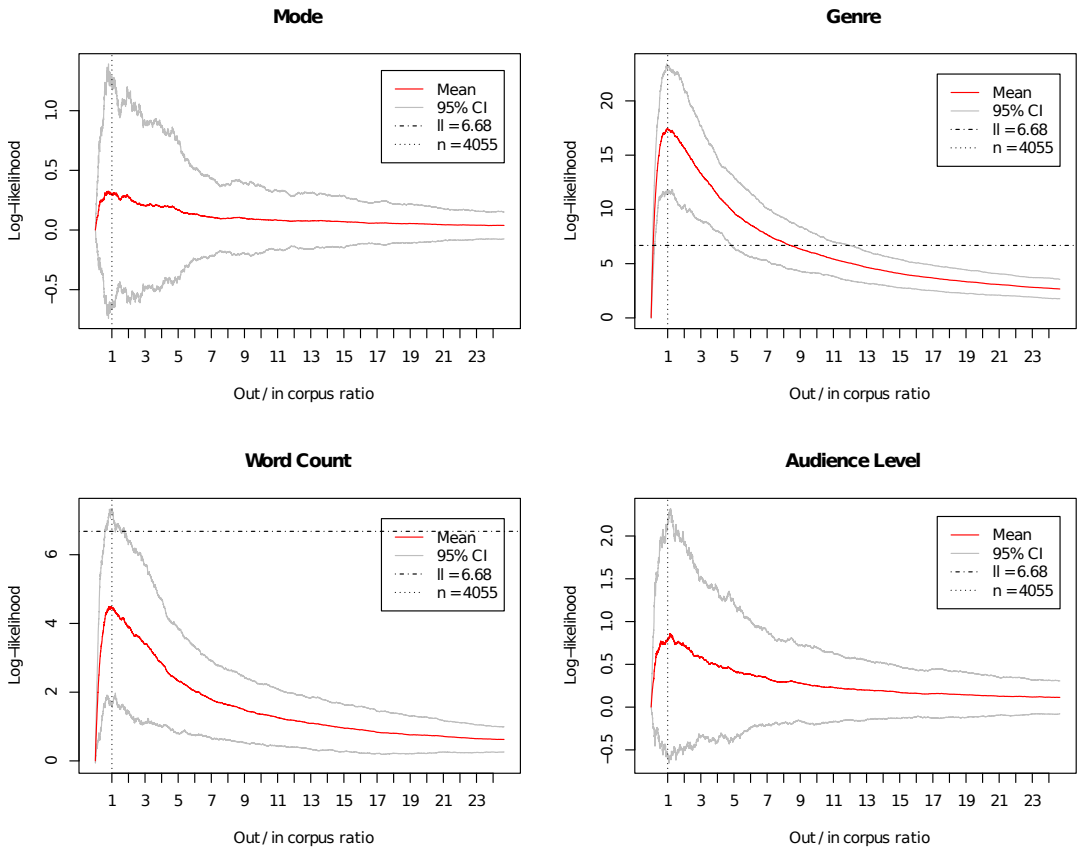


Figure 6.5: Log-likelihood scores and 95% confidence intervals for each metadata property during resampling (100 repetitions of 100,000 documents)

The two simpler distributions, 'Mode' and 'AudLvl', peak at levels well below the 95% threshold chosen for the log-likelihood statistic. In the case where the 'target' input distribution is defined by profiling an existing corpus, this implies that we could properly build a corpus *smaller* than the input (though unfortunately this is not rigorous without also knowing the relationship of the original input corpus to the population as it is itself generated from that sample). The relatively confident estimates for these parameters at a 1:1 input:output ratio backs up the rational expectation that large numbers of documents in few bins will be easy to replicate.

A marginally more complex distribution of metadata in the form of the 'Words' dimension shows, again, fast convergence, requiring an output distribution roughly twice the output to replicate with 97.5% confidence.

By far the most complex input data, 'Genre', takes much longer to converge, requiring a ratio of roughly 12:1. This result is largely a function of the low bin frequencies and the large proportion of bins that have values below the threshold at which log-likelihood may be relied upon to provide a moderately-conservative estimate.

The speed of convergence reflects the variability that is permitted in any subsequent retrieval process: fixing the 'Mode', for example, will lead to a corpus with known proportions of spoken and written text, but variability otherwise according to the population from which they are sourced. If a particularly large corpus is available with a desirable sample design, restricting

only one dimension merely refines the original sample[37].

The process of random resampling detailed here is in many ways uninteresting: after all, the experiment detailed above merely selects items based upon existing metadata. The value in the above technique lies in its mode of use:

- The distribution converged upon is arbitrary, and may be defined in the absence of an input corpus (or as a modification thereof). This allows relaxation of a corpus description's requirements to reduce the necessary oversampling ratio above, or the addition of specific metadata items where the source of documents is particularly easy to sample according to certain properties.
- Halting the resampling process at any stage, due to the random selection method used, leads to an unbiased output corpus (though it may well be imprecise).
- When starting with a manually-designed corpus (which has bin sizes only relevant to one another), convergence can only be achieved where the output corpus size is greater than or equal to that of the input corpus. Aside from fitting issues where bin sizes are small, corpora fit using the same distribution shape should converge at the same size (i.e. *not* relative to the input corpus size).

The relationship described by the log-likelihood scores of the two corpora describes the confidence one may have in the original, input, corpus: an input corpus with very few documents in each 'bin' is less powerful when used to answer questions using data in that bin, and so will require a greater number of documents in the output corpus to achieve confidence at the same effect size. This is a useful effect, as it ensures that samples are never output that are not in some way large enough to be confident about their distribution, thus enforcing a minimum sample size. Unfortunately, if this comparison is based on duplication of an existing sample (such as in this case) such an adjustment is non-rigorous as it is based upon the sample parameters, which have an unknown relationship to the population parameters they estimate. Essentially, whilst this is a heuristic indicating internal validity, it does not ensure external validity.

A notable property of the resampler is that, even if the confidence of the output corpus being 'converged' to a given confidence level is very low (i.e. the log-likelihood score is above a chosen threshold), the output corpus is unbiased relative to its input due to random selection. In addition, the convergence of a corpus with a simpler design than the source corpus will occur at a number of documents lower than that of the input corpus, making it possible to resample an existing corpus using a design with *simpler* metadata properties. In this case, the 'simplicity' of the design is defined by the information in the design distribution relative to the information content of the source corpus. For example, if our experiment did not care about genre, it would be possible to sample texts directly from word count and audience level, producing a corpus with fewer texts yet the same coverage of those two dimensions.

[37] Of course, such a situation is difficult if the corpus from which documents are not tagged with relevant metadata, something unlikely if that dimension was not already in the sample design

6.4 Performance of Retriever

The purpose of the resampling process above is to present well-defined document prototypes to a retrieval stage. There are many potential sources for document retrieval according to this prototype, including the original input corpus, a separate 'super-corpus', or, ideally a separate population (that is itself not a sample).

6.4.1 Method

The proof-of-concept evaluated here implements a mechanism similar to that of BootCaT[14], using a search engine to retrieve results based on key n-grams. This mechanism was selected due to the pertinence of its method to other corpus building efforts—there is no technical limitation imposed by the software tool, and users are free to select any other source of data or mechanism of retrieval.

The similarity of the retrieval system described here to BootCaT, along with the use of concrete prototype documents, allows us to inspect the 'measurement error' of using search engines to retrieve data based on key terms. This evaluation is, therefore, primarily an assessment of that bias and a preliminary assessment of its sources.

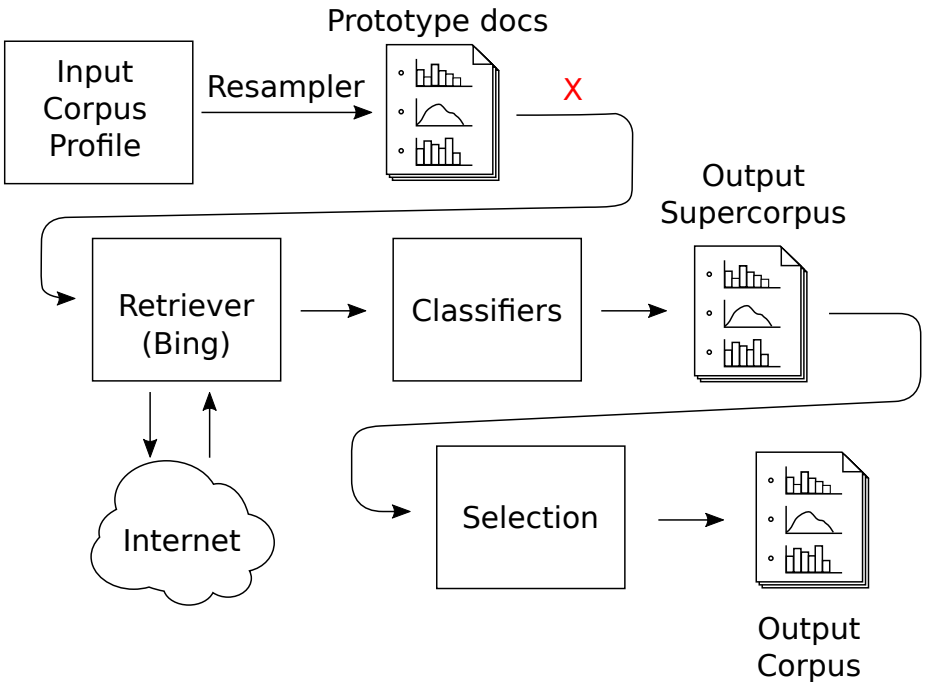


Figure 6.6: An overview of the retrieval mechanism used in the proof-of-concept implementation.

The retrieval process (as summarised in Figure 6.6) is performed iteratively using the components evaluated in the previous sections:

1. Sample a prototype document from the corpus profile;

2. Retrieve candidate links using the Microsoft Bing search engine[38];
3. Download, remove boilerplate (using JusText[139]), and classify each document using heuristics from the corpus profile;
4. Measure distance in the resulting vector space according to each heuristic's value.

For this data set, both the direct links from Bing and the first level of hyperlinks in each document were downloaded. Spidering was performed in order to ensure that the sampled area of metadata is a supersample of the desired distribution, as well as to maximise the chances of finding suitable points in other, uncontrolled, dimensions.

Documents with fewer than 100 words were discarded as they were unlikely to be classified accurately. This could be worked around by selecting 100-word samples from larger texts if necessary.

For this evaluation, the written portions of the BNC were used as the seed corpus. Keywords used for search were generated using log likelihood scores, and no cutoff was used: instead, the random selection algorithm was weighted by the resulting score.

The spoken and `W_Misc` categories were omitted for a number of reasons. Spoken data was omitted on the conjecture that it is difficult to find online [39], leading to a predictable gap in the resulting corpus. `W_misc` was omitted due to the need for an accurate classification step, and because keyword-based retrieval requires that said keywords are highly representative of a given genre. Both of these issues are potentially solvable in future work, yet lie outside the scope of this thesis.

The retriever must, as far as possible, return documents according to *all* prototype metadata dimensions, that is, it must perform the equivalent to an agglomerative query. This is a particular challenge for search engines due to their generality. Since the Bing API lacks tools to filter by any of the three dimensions used in this evaluation, the primary focus was on genre—word count and reading ease were both free to vary. Language was specified as `en_GB`.

6.4.2 Results

In total 72,540 documents were downloaded after 6,510 requests to Bing. After boilerplate removal using JusText[139] and discarding of short documents, 55,790 documents remained in the corpus.

This sample size is ≈ 21 times the input corpus. Estimating using graphical methods (as in Section 6.3, this should be sufficient to provide an appropriately distributed sample of genres. The worst case is that each Bing search returns very few documents with the desired genre, for example, if each search returns only a single document of interest then this ratio is reduced to $\approx 3n$.

Before inspecting the results from the retriever, the limitations of this method should be noted. Firstly, the ideal retrieval mechanism should target all dimensions of the prototype document

[38] https://datamarket.azure.com/dataset/bing/searchweb

[39] Retrieval of transcriptions may be facilitated by searching for genre-specific features in addition to keywords, and this form of specialism is a potential avenue for improving the accuracy of all genres.

provided, in this case Genre, reading ease, and word count). Bing does not index the latter of these two, leaving them to vary naturally.

This variation interacts with the genre itself, and so it is not possible to generalise from the observations here to the wider web. The distributions shown in Figures 6.7 and 6.8 are, however, indicative of the documents found via Bing when searching for BNC terms. Upholding the assumption that the written BNC represents general-purpose language use, this is an indication of users' exposure to documents online.

From the point of view of one sampling documents to form a supercorpus, it is desirable that the uncontrolled dimensions are widely and evenly dispersed across the vector space, such that any subsequent rejection sampling is simplified.

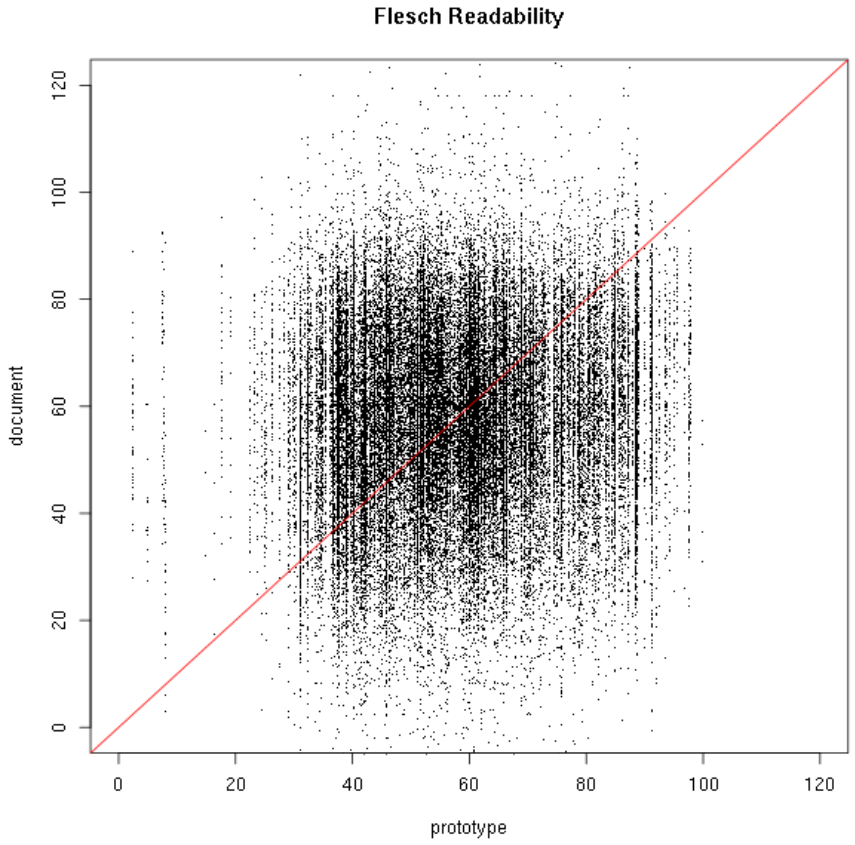


Figure 6.7: Prototype/document readability scores

Figure 6.7 shows the distribution of reading ease scores across the collected documents, the y axis showing the values as downloaded from the web. The red line is plotted with a gradient of 0.5, and shows the ideal zone where documents should lie if retrieved perfectly.

The readability dimension (Figure 6.7) is fairly well dispersed, indicating that there are no documents fundamentally unreachable using the search engine method. Though this dimension is the simplest, both the prototype and retrieved documents have similar standard deviations (16.2 and 17.9 respectively) and means (57.5 and 57.4 respectively). There is very little correlation between prototype and retrieved values: with a correlation coefficient of just 0.06. Generally, this indicates that most distributions of readability are 'covered' online, and the one used in the BNC

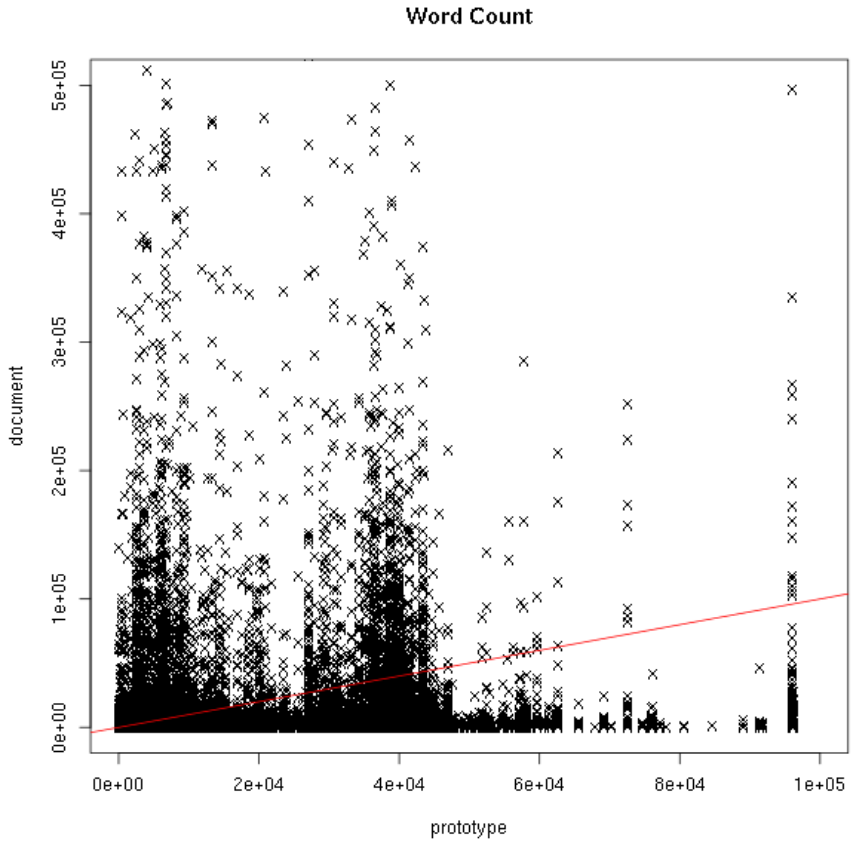


Figure 6.8: Prototype/document word counts

is easily represented.

Word counts are altogether more complex. The word count is distributed in the BNC in a far more complex manner (Figure 6.9), and the method of sampling can be reasonably expected to impact this more than readability. This shows that the distribution of word counts online (conditioned on genre) is far from that in the BNC: the line of intended fit failing to intersect with longer documents.

The distribution plots in Figure 6.9 display the obvious missing portion of the word count target distribution. This is a clear example of the bias of online documents, which tend to be paginated even where the content is particularly long (a trend that is exacerbated by advertising-based funding models). The bimodal distribution of the BNC's word counts can be attributed to interactions with 'domain': with `W_world_affairs` and `W_imaginative` forming the bulk of the longer texts.

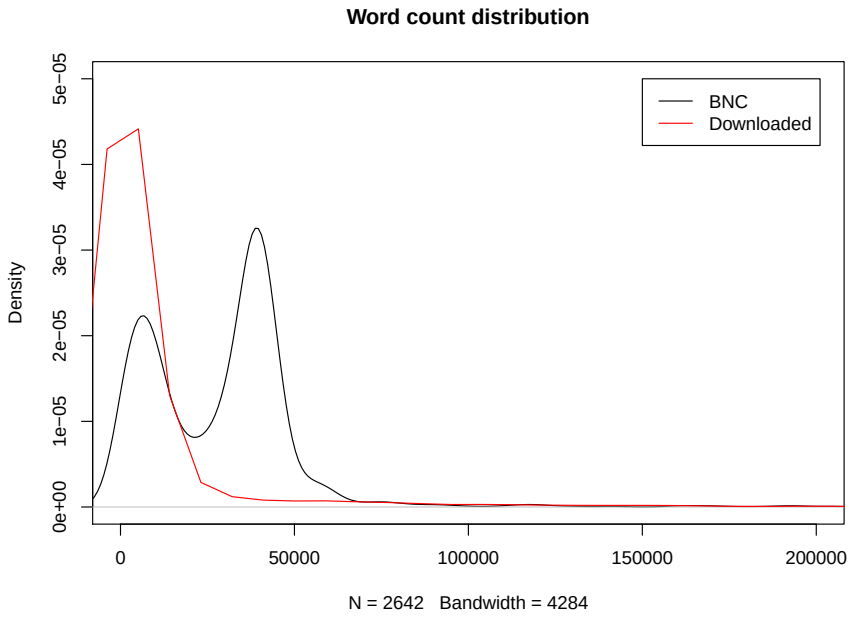


Figure 6.9: Distribution of document lengths in the written BNC and downloaded corpora

	Genre	Prototype	Downloaded	TP	TPR (%)
1	W_ac_humanities_arts	2180	1089	181	8.30
2	W_ac_medicine	987	1162	415	42.05
3	W_ac_nat_science	684	1258	160	23.39
4	W_ac_polit_law_edu	4424	1467	590	13.34
5	W_ac_soc_science	3112	1074	290	9.32
6	W_ac_tech_engin	466	1267	129	27.68
7	W_admin	195	1038	23	11.79
8	W_advert	856	4501	273	31.89
9	W_biography	2979	1736	221	7.42
10	W_commerce	2952	2058	464	15.72
11	W_email	46	882	3	6.52
12	W_essay_school	1	871	0	0
13	W_essay_univ	246	23	0	0
14	W_fict_poetry	815	460	42	5.15
15	W_fict_prose	10487	1844	888	8.47
16	W_hansard	161	114	39	24.22
17	W_institut_doc	572	1952	131	22.90
18	W_instructional	111	941	54	48.65
19	W_letters_personal	124	611	9	7.26
20	W_letters_prof	227	396	9	3.96
21	W_newsp_brdsht_nat_arts	764	3470	147	19.24
22	W_newsp_brdsht_nat_commerce	792	484	100	12.63
23	W_newsp_brdsht_nat_editorial	212	909	23	10.85
24	W_newsp_brdsht_nat_misc	1544	896	45	2.91
25	W_newsp_brdsht_nat_report	854	925	73	8.55
26	W_newsp_brdsht_nat_science	235	284	6	2.55
27	W_newsp_brdsht_nat_social	881	259	6	0.68
28	W_newsp_brdsht_nat_sports	639	503	97	15.18
29	W_newsp_other_arts	216	1345	49	22.69
30	W_newsp_other_commerce	91	466	18	19.78
31	W_newsp_other_report	756	933	52	6.88
32	W_newsp_other_science	327	196	3	0.92
33	W_newsp_other_social	904	705	41	4.54
34	W_newsp_other_sports	26	540	4	15.38
35	W_newsp_tabloid	45	482	2	4.44
36	W_news_script	479	121	3	0.63
37	W_non_ac_humanities_arts	1610	1665	193	11.99
38	W_non_ac_medicine	384	1605	142	36.98
39	W_non_ac_nat_science	1438	1534	258	17.94
40	W_non_ac_polit_law_edu	2891	1428	463	16.02
41	W_non_ac_soc_science	2210	1516	134	6.06
42	W_non_ac_tech_engin	2578	807	513	19.90
43	W_pop_lore	3371	6918	662	19.64
44	W_religion	838	1942	400	47.73

Table 6.4: Document counts by genre for prototypes and downloaded documents.

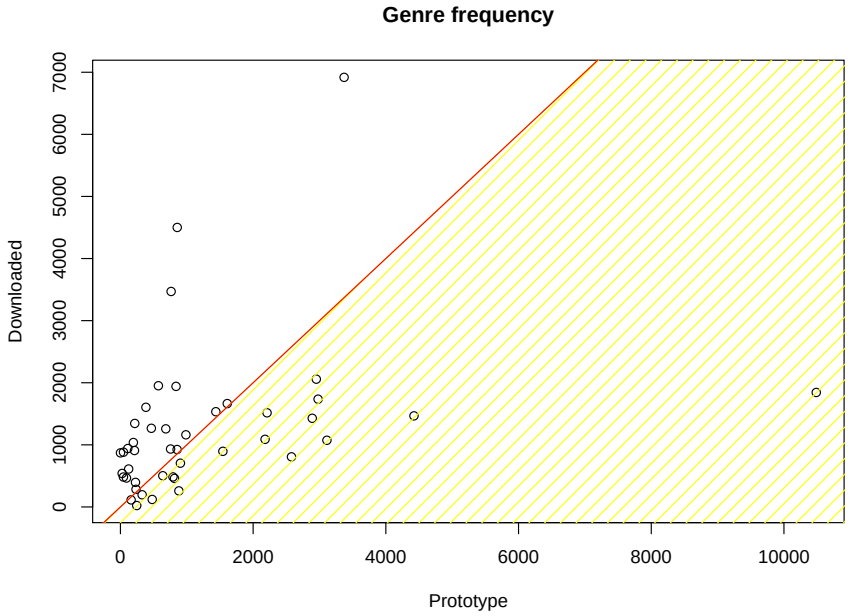


Figure 6.10: Prototype genres frequencies and resulting downloaded document frequencies. Shaded area indicates 'over-downloaded' categories.

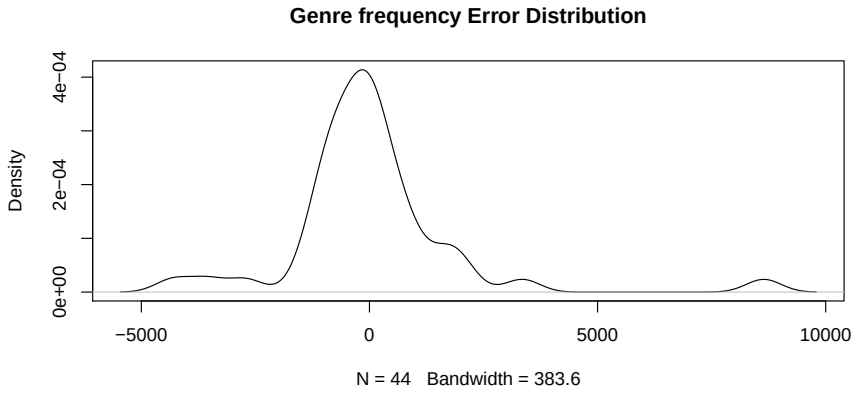


Figure 6.11: The distribution of differences in genre frequency between prototype and downloaded documents.

Genre was the only parameter explicitly controlled-for by the search engine retrieval method. Precise counts for each of the prototype and retrieved genres are provided in Table 6.4.

Figure 6.10 shows the prototype vs. downloaded genre frequencies. Those genres above the line are under-represented in the downloaded documents, relative to the prototype documents[40]. When selecting a full corpus using the search engine lookup method, differences between the prototype and downloaded document classes should be minimal—when selecting a supercorpus, care must be taken to ensure that items in the top-left of the plot have sufficient frequencies for the intended purpose.

[40]This measure is distinct from precision/recall in that it does not care which individual documents account for the frequencies, only that the search has retrieved appropriate proportions of each

Evidence for the 'directedness' of genre-based retrieval using n-grams is given by the dispersion of data around the red line. Rank correlation between prototype and retrieved documents is 0.45, and a Mann-Whitney test is not significant ($U = 786.5; p \approx 0.13 > 0.05$), indicating that errors in retrieval are loosely centred around the prototype's frequencies. This is reinforced by inspecting the density plot for differences between category frequencies, shown in Figure 6.11.

Treating the searching process as a classifier allows us to examine the expectation of retrieving a genre successfully. 7,355 documents were returned with the desired genre, leading to a true positive rate of 13.2%. This implies retrieval results close to the worst case of one relevant document per search, and explains the low correlation between predicted and resultant class types. Note that the number of useful documents is above this thanks to re-examination of those previously returned.

Per-class retrieval rates are displayed in the right-most column of Table 6.4. There is correlation between the size of a genre and its true positive rate ($\rho = 0.29$), and inspection of the table indicates that performance is highly variable. Particularly low retrieval rates (below 5%) were reported for ten categories. Many of these are very 'general', defined by their context rather than topic: such topics both resist keyword-based retrieval using search engines, and make classification itself difficult once retrieved. The notable exceptions are the `_science` and `_social` news categories, retrieval rates of which are likely to be affected by the anachronistic nature of the keyword source (the BNC). Notably, though retrieval rates were low for these categories, documents were available in all cases, they were simply not the result of that particular search iteration.

Though it is not possible to separate classifier error from retrieval error without construction of a gold standard corpus (something that lies beyond the scope of this work), earlier evaluations of the classifier used (Section 6.2.3) show that classifier error is not strongly biased. This allows us to conclude that evidence of significant bias in the resultant data is largely a product of the searching process, rather than document classification.

Finally, keywords were computed relative to the original input corpus using log likelihood. These offer insight into the specific biases of the data source and a handle on qualitative differences between sampling processes.

The top 200 keywords[41] were inspected for the whole supercorpus relative to the seed corpus. Overrepresented tokens were scored separately to under-represented ones in accordance with Baron et. al.[13], and the lists were free-coded for prominent themes. The 20 most over- and under-represented terms are displayed in Table 6.5. The full lists are included in Appendix F.

Overrepresented terms may be grouped into a number of categories. The first of these is technical terms: those that are side-effects of the tokeniser or otherwise are related specifically to the web. This includes features such as protocols (*HTTP*), and features of URLs (*co*, *org*), as well as terms such as *email* and *posted* which have simply become more prevalent since the BNC was produced. It's difficult to say whether or not a more recent seed corpus would eliminate

[41] I recognise that nspection of the top n keywords is not a hugely sound methodology, but it should prove indicative for these purposes

Over-sampled				Under-sampled			
term	**original**	**retrieved**	**loglik**	**term**	**original**	**retrieved**	**loglik**
p	29217	8839489	1977900.5	she'd	7338	4050	-17407.3
s	17778	1903037	357245.2	darlington	5559	1521	-16657.0
she	290860	916731	136365.1	ec	6091	5531	-11220.7
gt	111	535127	134335.1	solicitors	2135	1086	-5237.0
com	172	465891	116044.3	swindon	2182	1409	-4825.7
www	2	448772	113981.6	kinnock	1375	264	-4466.7
amp	668	432633	102780.7	theda	838	32	-3297.3
lt	792	425436	99921.8	nizan	756	6	-3146.1
had	359670	1439973	94877.1	cnaa	742	1	-3140.1
h	8094	555374	90772.7	guil	755	43	-2886.5
t	9078	557616	87217.1	maastricht	1165	593	-2856.9
her	277189	1070672	80897.8	oesophageal	957	279	-2820.0
2015	25	293434	74145.0	lexical	1106	514	-2809.1
2014	16	254997	64517.4	robyn	1200	707	-2767.5
he	523177	2629585	59109.8	lamont	1184	708	-2712.6
was	722425	3866921	58399.0	risc	1013	416	-2689.4
cent	34353	44410	49457.4	knitting	1509	1463	-2668.9
the	5021789	33346038	47673.9	athelstan	1060	516	-2645.3
be	524709	2771036	45819.9	privatisation	1104	665	-2521.1
it	727487	4058706	45306.4	leapor	626	18	-2502.2

Table 6.5: The 20 most over- and under-represented terms relative to the BNC.

most of these from the keyword list.

Temporal features dominate the rest of the list, with many terms related to date or proper nouns (*Obama*, *Nike*). Dates feature prominently: *2001-2012* are all mentioned, along with *1991* and *1988*. Genetics is also mentioned (*gene*, *mutation*), along with education (*university*, *pupils*) and government (*government*, *council*).

The remaining terms are pronouns or function words. These, along with the large number of quantities and references to time/date imply consistency with other WaC datasets, that is, an overrepresentation of news online[169].

This central theme is interesting to contrast against the BNC, which contains a large quantity of news material already, indicating that news genres are even more prevalent online, or that news media (or perhaps just online news) has become more extreme (that is, less similar to other written genres) than the bulk of the BNC data.

Features that are significantly under-represented predominantly fall into the category of proper nouns, and these are mostly political and from the UK (*Kinnock*, *Gummer*, *Heseltine*). Since the search engine was instructed to search the 'GB' marketplace, many of these are most likely explained by temporal effects—the prevalence of these keywords also points at the strength of news within the BNC, along with places and organisations that featured in the news at the time (*TUC*, *Maastricht*, *DTI*).

There is also a strong theme of legal terms (*offeror*, *solicitor*, *arbitrage*), which points at a clear under-representation of legal documents returned by search engines. This is likely a technical limitation of the retriever, which avoids files in PDF format.

Some technical terms are also presented in the keyword list due to the speed of progress in that field: *80486* and *microcomputer* both showing their age.

Finally, there are a number of gastric terms (*oesophegeal*, *pylori*, *duodenal*), which point to a very specific set of documents missing in the returned data. These terms are from an academic context that, like legal documents, is under-represented for technical reasons, however, it is also likely that the specificity of this category indicates an over-representation within the BNC to some degree.

The relative obscurity of the under-represented terms and the predictable, largely temporal, nature of the over-represented ones is particularly encouraging, indicating that the web is a suitable source for many general-purpose (or current-affairs-focused) applications.

6.5 Discussion

The evaluations above form a white-box exploration of the characteristics of components in a context similar to many real-world demands. Overall, performance is encouraging: though there are significant challenges to the accuracy of classification and retrieval, combination of the three main stages is able to retrieve a corpus with only minor changes from its target input.

Firstly, the performance of classifiers when operating on text must be established relative to the research question. In this case, that was particularly tricky for a number of practical reasons which shadow the classical manual corpus construction issues:

- Linear separability of dimensions such as readability was insufficient to create a classifier with good precision/recall scores;
- The complexity of some sampling criteria requires the use of black-box models that are themselves unpredictable, and require training using auxiliary data.

Both of these are common cases when dealing with 'human', research-aligned sampling criteria, and both are rightly recognised as key areas for improvement in text processing. Any deficiencies in real-world performance of this classification process must be assessed on a case-by-case basis: for certain sampling criteria, classification is made trivial by the existence of metadata (for example the recipient fields in emails) or existing mechanisms online for directed retrieval.

This limited classification capacity comes with a silver lining, however: unlike other bootstrapping approaches, it is possible to separate and independently evaluate the performance of any classifier used. In turn this makes it possible to re-use others' research in this area and make incremental improvements over time. Secondly, classification accuracy is improved by application to well-specified problems of limited scope: a priority that aligns well with the core values of a good sampling scheme. The evaluation shown above indicates that error in the models chosen here are largely unbiased, meaning they only minimally impact on the utility of a supercorpus.

Examination of the BNC categories used here also reveals the loose nature of classification therein: many categories are defined according to fairly fluid concepts of varying specificity, and many overlap significantly. This is evidenced by the relative performance of classifiers built

with minor changes to their classes, and by the keywords used by the Bayesian models. Further work is required to define a useful yet strictly-defined set of genre categories—this may be best performed by those creating corpora if no community-wide agreement can be made on salient categorisation schemes.

The bootstrapping process, separated into a process of building a prototype corpus, has been shown to converge within a timeframe that is compatible with automated retrieval.

For the BNC example outlined here, roughly 12 times the input corpus size is required to converge with 95% confidence. This coefficient is based on the complexity of the input corpus, making the technique better suited to simpler corpora, or those with lower variance parameter spaces (such as special-purpose corpora, or corpora built primarily around small numbers of parameters).

The more complex each document, the more difficult it is to retrieve. This relationship is unfortunate, as it is more likely that a greater number of documents will be required in order to construct a corpus.

One way around this is to sample using a different sampling unit. The method presented here is capable of word-, paragraph-, and sentence-level retrieval, something that also brings significant methodological benefits by reducing or eliminating the need for statistical corrections to compensate for the dependent nature of words.

Automated retrieval is the part of the method with the greatest number of practical challenges. The method used here is deliberately comparable to the bootstrapping approach used by BootCaT[14], in part to use the prototype corpus approach for an evaluation of search engine retrieval methods.

The evaluation here focuses on genre-based retrieval. This is of particular salience to many linguistic research questions, and is well aligned to search engine indexing behaviour—an ideal automated retriever would go far beyond this to use document metadata and different sources of text (such as academic repositories).

Keyword-based search engine retrieval is shown, in the case of the BNC, to have roughly central error—this is encouraging, though it is unclear how well this holds for other corpora. Of particular interest is the application of the method to special-purpose corpora, which are (by design) strongly biased. It should be possible to use the evaluation methods presented here to gradually reduce this retrieval error, refining search based techniques in the process.

We have also observed the systematic over-sampling of certain genres. This implies that automated retrieval must advance in order to sample many corpus designs, or that it must be complemented by manual methods where documents are difficult to locate. The exact thresholds of 'difficulty' there must be determined on operational grounds of cost and time: for large-scale corpus building operations, a partially manual approach may be the best method of ensuring quality whilst still leveraging the benefits of the quantitative profiling and resampling approaches.

This method provides a mechanism for inspecting and evaluating these biases, and repetition with different corpus designs would reveal how pervasive they are. This is a key difference

between the method presented here and existing web-based corpus retrieval, which is largely monolithic.

Ultimately, this evaluation has shown the method to be largely suitable for constructing general purpose corpora online, in a manner similar to existing tools. Insights gleaned from the examination of each component indicate that it is also suitable for certain special-purpose corpora, and that its output may be generalisable to subsamples of the BNC.

Quite how far it is possible to 'push the envelope' of genre distribution is unclear, though this must be established through repetition and comparison to many known targets. Currently, the ability to do that is hampered by the state of classification and retrieval technology, though neither of those are intractably difficult to improve, and both yield value to many users.

Ultimately, even if a semi-manual document selection method is used for the final stage, this method is still significantly faster and easier than full manual selection. The BNC-based corpus sampled here, for example, was retrieved automatically over the course of roughly a week, with minimal user interaction.

The primary value of this method is the decoupling of retrieval and resampling mechanisms from the underlying data. This allows proportions of data to be designed manually without providing a 'seed corpus', meaning it is possible to construct corpora automatically simply from a sample design.

6.6 Summary

This chapter, and the one preceeding it, have specified and tested a generalisation of many web based corpus building systems. The intent is that this method is split into modules which are easily tested in isolation and which represent separable concerns within the corpus-building process.

This evaluation has focused on a common-case of constructing a corpus based on an existing set of metadata, following the stage of building a transferrable profile for the BNC, resampling a set of prototype documents, and then retrieving a corpus using the Bing search engine. A white-box approach provides some insight into the transferrability and utility of each stage in a real-world context, and the selection of corpus and retrieval methods (in line with many existing methods) provides some transferrability.

The classification accuracy for the parameters chosen in the BNC was variable: the task of classifying genres is a challenge for current machine learning methods, and this seems set to continue, especially where fine-grained categories are used. This challenge alone implies that the method is better suited to special-purpose corpora, or problems that do not require controlling for genre[42].

Bootstrapping to retrieve prototype documents is predictable, and indicates that corpus sizes are not intractably large for modestly complex input distributions. The complexity (and thus

[42]Though doubtless very few such problems exist.

the required number of documents) increases massively with the addition of new dimensions, meaning again that simpler designs are better suited to quick retrieval. Manual simplification of the distribution (such as binning continuous distributions like word counts) offers a way to control this manually whilst managing losses in accuracy.

The final, retrieval, stage is particularly challenging. The retrieval mechanism here exhibits similar properties to that used in BootCaT, mainly the over-representation of news sources, and this seems unlikely to change for other inputs. The prototype-based approach, however, offers a mechanism for further investigation into the bias of the retrieval mechanism, and there is significant further work necessary to evaluate different methods. Investigation of retrieval errors, repeated with different study designs, could eventually be used to build up a picture of search engines' suitability for a given task.

For the BNC-based task outlined here, the final output corpus was largely a suitable supercorpus. The next stage for any user of this corpus would be to discard any documents with more error than is deemed damaging for a given study design, something I am unable to do here without assuming a large number of theoretical conditions. Areas under-represented in the retrieved data are largely technical—these could be retrieved by sourcing data from places other than a general-purpose search engine.

Ultimately, this method was able to produce a BNC-like corpus quickly and with minimal end-user interaction, in a manner that is described fully as a single, human-readable file. This opens up opportunities to repeat sampling runs and produce sample designs without having to rely on existing corpora as a source of 'seed' terms, and offers a way to evaluate the corpus construction process with known margins for error. The externally-defined nature of the corpus definition also relieves users of many issues surrounding dissemination of corpora, easing documentation, licensing, and bandwidth requirements.

Chapter 7

Conclusions & Review

This thesis has examined the problem of improving corpus sampling from a number of perspectives. This chapter will serve to summarise these, before offering an overview of significant areas for further research.

Chapter 2 provided an overview of current corpus construction techniques from two main perspectives: the former is that of corpus linguistics itself, which has debated issues of representativeness largely from the point of view of improving existing corpora. The latter is that of sampling in other fields (largely the social sciences, still), which is able to provide a more theoretical framework.

Comparisons between these two yield a number of issues with current corpora, largely focusing on their population definition, and the relationship between expert opinion and demographic frequency. New recording technologies, and sources of data such as the web, offer ways to make some progress on both of these.

Chapter 3 presents a motivating study of link rot in open-source corpora, finding large variations in document half-life using a URL-seeking method. This illustrates that temporal effects have a significant impact on the availability of data, and that such effects vary according to source. Much literature exists that has related these to layout features (such as the role of navigation and landing pages), yet the impact this has upon linguistic content is less well understood.

In order to make such investigation possible, a cohort sampling approach is necessary, performed at a high resolution over a long period of time. The LWAC tool automates this process, providing a mechanism to construct large-scale corpora that are indexed by time as well as URL. Key to the utility of such a tool is its performance, which was shown to be adequate to construct corpora in the millions of documents, sampled daily.

This wealth of data enables analysis not only of 'simple' availability, but also changes through time due to editorial processes or site redesigns. This may allow sociolinguistic analysis surrounding specific events, or a more general picture of long-term language change through time. It is hoped that the latter of these can be used as auxiliary data to inform stratification of future general-purpose corpora (or at least their web-based components).

Chapter 4 presents a case study detailing a short-term linguistic census of a single subject. The sample design used therein is a 'narrow and deep' design, orthogonal to the 'broad and shallow' coverage of the large national corpora. This makes it particularly useful for judging rationally how well such data apply to individuals, and offers a way to retrieve metadata not found using conventional data gathering methods.

Such a detailed sample also provides a mechanism for reasoning about problems of corpus size: the data observed in just two weeks for a single subject yielded almost a million words. Assuming significant inter-person variation for a given linguistic feature, this implies that a population of 60 million people cannot be well represented by a corpus of just a few million words. Quantitative knowledge of inter-person variation is needed to operationalise this, but such data could be gathered using some of the automation methods presented.

Such methods were largely taken from lifelogging, a field chosen due to its focus on unobtrusive, best-effort, and wide-ranging sampling techniques. These proved to be significantly more troublesome than expected: though automation was assistive, much manual correction was necessary to operationalise the data, including the need for a human-interest model to correct summary statistics that were gathered at a low resolution.

Chapters 5 and 6 cover a method for 'profiling' a corpus purely based on independent variable specification. This profile has many potential uses, both as human documentation for a corpus and as a machine-readable descriptor, and constitutes a way to gradually refine corpus designs over time.

This system is one designed to extend and generalise the approach taken by BootCaT in order to make its selection criteria easier to operationalise. It is designed to be modular, allowing those building a corpus to re-use algorithms and taxonomies in a simple manner: essentially, the aim is that the tool should be working using the same concepts and distinctions as the desired sample design.

Monte Carlo techniques were used as a method for operationalising the profile, something that was shown to achieve convergence to the desired distribution within a tractable period for a BNC-like corpus of written text. The classifiers required to identify documents for final selection were shown to be workable, but their performance was borderline for the more complex dimensions such as genre—this is in part due to the ambiguity in the test data's taxonomy, and in part due to the difficulty of large-scale classification tasks. Such issues are of wide utility, however, implying that improvement in this area will not require specific study.

Retrieval of documents according to the prototypes was shown to be possibly the biggest challenge. This is in part due to the classic corpus sampling problems of being unable to locate documents in a uniform and reliable manner, and in part due to the biases seen online. A number of methods were presented to work around these issues, such as manual selection or use of manually-curated web indices, and evaluation of these remains as further work.

To revisit the problem of representativeness, this thesis has taken the view that any such question must be framed in terms of a larger study design. This is not to say that general-

purpose corpora are impossible to create: sampling larger populations is such a challenging and resource-intensive endeavour that it will most likely always have to be offloaded onto a third party.

The aim, then, is to provide a general-purpose corpus that is uniformly representative for a wide, and well-specified, set of research questions. This corpus should be sampled using random sampling techniques, and this thesis' tools have been designed to inform stratification efforts with this in mind. Such a corpus should also be unambiguously and comprehensively documented: assumptions should be made explicit in corpus documentation so that they may be compared against those made by researchers, and the limitations of the intended use of a corpus should be stated in quantitative terms.

The contributions within this thesis have been targeted to provide some insight into the underlying population that is not possible using current techniques

In accordance with **RQ1**, a number of sampling methods have been rendered accessible to corpus linguistics by the tools presented here. Longitudinal sampling is now available as a strictly regulated, cohort design with many quantitative regressors, and quantified, operationalisable stratification is possible through use of the corpus profiling tools in Chapters 5 and 6.

WaC methods are used in two ways: both as a way of constructing corpora entirely, and as a mechanism to inform other corpora.

The former of these, the focus of **RQ2**, is addressed by Chapters 3 and 4, both of which use the web as an easy source of information to construct corpora that are able to inform design choices for more conventional approaches.

Chapters 5 and 6 focus on being able to operationalise any insights from this, providing an opportunity to incrementally improve sampling for general-purpose corpora (the focus of **RQ3**).

Specifically, the contributions of this thesis are:

- Tools for sampling documents from the web according to a cohort sampling, longitudinal design;
- A method for sampling 'personal corpora' as language is used, and a discussion of the implications this has for conventional corpus building;
- A method for resampling corpus data stratified according to user-selected linguistic variables, and representing the distribution therefrom;
- Software tools for automatically reconstructing corpora from designs described using said profiles.

7.1 Further Work

As much of this thesis has been exploratory, there is a wealth of further work necessary to maximise the value therein. Whereas each chapter contains numerous notes in context, this is a high-level summary of the major avenues of inquiry.

7.1.1 Large-Scale Longitudinal Document Attrition Studies

A large-scale longitudinal sample of WaC sources is one obvious extension to the work in Chapter 3[43].

Such a study would be able to answer questions not only on the rate of decay (as has been addressed elsewhere), but also the shape of the decay curve, hitherto presumed to be exponential for most types of data.

Further to this, there are many web page changes that do not constitute 'link rot' in that the original information may still be present, but are still interesting for various linguistic and social questions, for example how often documents are changed in response to news events, or have their boilerplate changed in order to affect the context in which documents are presented.

Ultimately, developing a linguistic understanding of change through time will require significant work, however, a single large sample could provide useful information to those seeking to use large existing web or monitor corpora.

7.1.2 LWAC Distribution

LWAC's distributed design lends it particular powers in distinguishing web access issues from a geographical perspective. With the low price of virtual machine rent, a network of LWAC clients could easily be constructed to access a number of websites from differing locations.

Such a configuration would reveal geographical changes made by site administrators for reasons of editorial control, censorship, and interference by service providers (Phorm[33], for example). This would open an avenue for investigation of a number of social factors, particularly as the (time-series) results could be compared and correlated with many real-world regressors, such as the incidence of political events. Such approaches are already being widely used in social media analysis[1, 27], though the high number of societal covariates renders such studies difficult to validate.

7.1.3 Slimmed-down Personal Corpora

A natural progression of the work in Chapter 4 is the repetition with more subjects. This is troublesome in part due to the in-depth coverage of language sampled, something that is likely to be less than critical for many research questions (such as those primarily concerned with 'standard' day-to-day interactions, or even simply those using a computer).

There are many mechanisms which could be used to establish a provisional model of demographic linguistic variation based on partial sampling of each subject, perhaps elective using smartphone technology, down to the level of questionnaires. This dataset could be added to gradually, and used to guide large diachronic sampling efforts as a form of auxiliary data.

This would require work to formalise and proceduralise the data gathering methods presented, in order to discount those with the least 'effort efficiency' and maximise automation. The

[43] Indeed, one was underway, but due to technical failures the results were not ready for publication here.

continuing ubiquity of smartphone technology offers a vehicle through which to distribute such methods and collect resulting data.

Operationalising the results of such sampling is also an open question, though such work would largely require the application of existing stratification and re-weighting techniques. This would allow an experimenter to select a small sample demographic, and then use a general-purpose corpus to augment it with larger quantities of linguistic information in a representative manner. In order for these methods to be useful at a large scale, they should be informed by further work on power analysis, which is currently in its infancy in corpus linguistics.

7.1.4 Augmented Personal Corpora

Metadata distributions from personal corpus retrieval may be used as input for ImputeCaT sampling techniques. Further, a corpus designed according to this method would be simpler to distribute, and would be scalable without requiring long-term data gathering from subjects.

As corpus proportions are encapsulated in the corpus description, they may be inspected and modified quantitatively. This allows for incorporation of data from other profiles, in order to balance data demographically, or from entirely exogenous sources (for example, document publication dates as taken from LWAC). Control over these may be viable approaches to retargeting a corpus for a given task, or updating it to suit changes in population.

One of the major challenges here lies in improving the classification tools used within ImputeCaT so that metadata on a personal corpus may be accurately retrieved. This is perhaps simpler for web history data and documents of digital origin, though manual retrieval of documents may be able to expand this to other published work.

7.1.5 Improved ImputeCaT Modules

The performance of the ImputeCaT corpus profile retrieval tool is currently constrained primarily by two factors:

- The accuracy of classifiers used to identify document distributions and class membership during profile generation and candidate document selection;
- The specificity of retrieval methods.

The former of these is the subject of much ongoing research, and can be expected to progress gradually due to its utility in many areas. This is still challenging, however, due to the use of web data. Appendix E shows a preliminary evaluation of the BNC45 classifier used in this thesis as applied to web data. The design of ImputeCaT is such that this may be leveraged easily to provide continuing accuracy improvements.

The latter is more challenging: though ultimately corpora can be retrieved manually (incurring only the error seen in inter-annotator scores for a given taxonomy[161]), the scalability of automated corpus construction is born largely of its ability to retrieve documents online in an

unsupervised manner. This is currently limited by the search-engine-based retrieval mechanism demonstrated here.

Many options exist for this, such as the use of web directories, supercorpora, semi-supervised topic modelling, or crowdsourcing. Each of these should be evaluated relative to typical corpus construction tasks in order to determine its suitability for given study designs: indeed, it is likely that some exhibit biases such that they are unable to replicate some samples.

7.1.6 Bias Estimation

Finally, the convergeance-based design of profile-guided retrieval offers the first few avenues through which to explore the bias of sources. Currently, bias in retrieval methods is conflated with classification error, something that is also far from negligable.

The approach of seeking a series of known document properties, however, leads to the obvious method of simply measuring how 'difficult'[44] it is to retrieve a document given certain parameters. Differences between this difficulty-of-retrieval metric and an equivalent measured based on human performance should yield biases in online sources.

Such a measure would be highly dependent on many unstable variables, however, such as the retrieval methods used, sample design, and temporal/contextual factors. Nonetheless, this offers a real-world take on representativeness without reliance on variance and homogeniety measures.

As ever, these methods are dependent upon finding a human analogue task that acts as a well-agreed-upon baseline: ultimately, without such a research question, the question of representativeness is unanswerable.

C'est fini!

[44] Ideally measuring difficulty for a typical human.

Bibliography

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using Twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707, April 2011.

[2] B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised historical text: Does it have to be mediocre. In *Proceedings of KONVENS*, pages 401–409, 2012.

[3] A. L. Allen. Dredging up the past: Lifelogging, memory, and surveillance. *The University of Chicago Law Review*, pages 47–74, 2008.

[4] L. Anthony. Antconc (version 3.2. 2)[computer software]. *Tokyo, Japan: Waseda University*, 2011.

[5] U. D. Archive et al. UK data archive. http://www.data-archive.ac.uk/. Accessed 2015-08-19.

[6] G. Aston. Text categories and corpus users: A response to David Lee. *Language learning and technology*, 5(3):73–76, 2001.

[7] S. Atkins, J. Clear, and N. Ostler. Corpus design criteria. *Literary and linguistic computing*, 7(1):1–16, 1992.

[8] R. H. Baayen and F. J. Tweedie. Sample-size invariance of LNRE model parameters: Problems and opportunities. *Journal of Quantitative Linguistics*, 5(3):145–154, 1998.

[9] P. Baker. The BE06 corpus of British English and recent language change. *International journal of corpus linguistics*, 14(3):312–337, 2009.

[10] J. Bar-Ilan and B. Peritz. The life span of a specific topic on the web. *Scientometrics*, 46(3):371–382, 1999.

[11] A. Barbaresi. Finding viable seed URLs for web corpora: a scouting approach and comparative study of available sources. *EACL 2014*, 2014.

[12] V. Barnett. *Sample survey principles and methods*. Edward Arnold London, 1991.

[13] A. Baron, P. Rayson, and D. Archer. Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67, 2009.

[14] M. Baroni and S. Bernardini. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC*, volume 4. Citeseer, 2004.

[15] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009.

[16] M. Baroni and M. Ueyama. Building general-and special-purpose corpora by web crawling. In *The 13th NIJL international symposium, language corpora: Their compilation and application*, pages 31–40, 2006.

[17] L. Bauer. *Manual of information to accompany the Wellington corpus of written New Zealand English*. Department of Linguistics, Victoria University of Wellington, 1993.

[18] V. Belikov, N. Kopylov, A. Piperski, V. Selegey, and S. Sharoff. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In *Proceedings of Web as Corpus Workshop (WAC-8), Lancaster*, 2013.

[19] G. R. Bennett. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press Michigan, 2010.

[20] D. Biber. Variation in speech and writing. *Chicago: University of Chicago*, 1988.

[21] D. Biber. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2):133–163, 1992.

[22] D. Biber. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257, 1993.

[23] D. Biber. Using register-diversified corpora for general language studies. *Computational linguistics*, 19(2):219–241, 1993.

[24] Z. W. Birnbaum and M. G. Sirken. Bias due to non-availability in sampling surveys. *Journal of the American Statistical Association*, 45(249):98–111, 1950.

[25] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, Apr. 2012.

[26] J. M. Blum. Early English Books Online. *Reference Reviews*, 16(4):5–5, 2002.

[27] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[28] L. Bowker and J. Pearson. *Working with specialized language: a practical guide to using corpora*. Routledge, 2002.

[29] G. E. P. Box and M. E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 06 1958.

[30] L. Burnard. Users reference guide British National Corpus version 1.0. 1995.

[31] L. Burnard. Users' reference guide for the British National Corpus, 1995.

[32] V. Bush. As we may think. *SIGPC Note.*, 1(4):36–44, Apr. 1979.

[33] R. Clayton. The phorm 'webwise'system. *Light Blue Touchpaper*, 4, 2008.

[34] J. Cohen. *Statistical power analysis for the behavioral sciences (rev)*. Lawrence Erlbaum Associates, Inc, 1977.

[35] P. Collins and P. Peters. The Australian corpus project. *Kytö et al*, pages 103–121, 1988.

[36] M. Davies. The 385+ million word corpus of contemporary American English: Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190, 2009-06-01T00:00:00.

[37] M. Davies. The corpus of contemporary American English as the first reliable monitor corpus of english. *Literary and linguistic computing*, page fqq018, 2010.

[38] K. J. Devers and R. M. Frankel. Study design in qualitative research—2: Sampling and data collection strategies. *Education for health*, 13(2):263–271, 2000.

[39] J.-P. Dittrich, M. Salles, and S. Karaksashian. iMeMex: A platform for personal dataspace management. In *SIGIR PIM Workshop*, pages 22–29, 2006.

[40] S. Dollinger. Oh Canada! towards the corpus of early Ontario English. *Language and Computers*, 55(1):7–25, 2006.

[41] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79. ACM, 2003.

[42] H. S. Eaton. *Semantic frequency list for English, French, German, and Spanish: A correlation of the first six thousand words in four single-language frequency lists*. University of Chicago Press, 1940.

[43] D. P. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, CARPE'04, pages 39–47, New York, NY, USA, 2004. ACM.

[44] P. D. Ellis. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press, 2010.

[45] S. Evert. A simple LNRE model for random character sequences. In *Proceedings of JADT*, volume 2004, 2004.

[46] S. Evert. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2):177–190, 2006.

[47] S. Evert and M. Baroni. zipfr: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 29–32. Association for Computational Linguistics, 2007.

[48] F. Farrokhi and A. Mahmoudi-Hamidabad. Rethinking convenience sampling: Defining quality criteria. *Theory and practice in language studies*, 2(4):784–792, 2012.

[49] A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[50] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[51] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[52] W. Francis and H. Kucera. Brown corpus of standard american english. *Brown University, Providence, RI*, 1961.

[53] R. B. Fuller. R. Buckminster Fuller papers. M1090. Dept. of Special Collections, Stanford University Libraries, Stanford, Calif. Available at http://www.oac.cdlib.org/findaid/ark:/13030/tf109n9832/.

[54] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: A personal database for everything. *Commun. ACM*, 49(1):88–95, Jan. 2006.

[55] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: fulfilling the Memex vision. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238. ACM, 2002.

[56] J. Gemmell and Z. Wang. Clean living: Eliminating near-duplicates in lifetime personal storage. Technical report, Tech. rep., Microsoft Research Report No. MSR-TR-2006-30, 2006.

[57] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[58] B. G. Glaser. The constant comparative method of qualitative analysis. *Social Problems*, 12(4):pp. 436–445, 1965.

[59] B. G. Glaser. *Emergence vs forcing: Basics of grounded theory analysis*. Sociology Press, 1992.

[60] Y. Goldberg and J. Orwant. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, pages 241–247, 2013.

[61] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[62] S. Greenbaum and G. Nelson. The international corpus of English (ICE) project. *World Englishes*, 15(1):3–15, 1996.

[63] S. Greenbaum and J. Svartvik. *The London-Lund Corpus of Spoken English*. Lund University Press, 1990.

[64] S. T. Gries. Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2):109–151, 2006.

[65] S. T. Gries. Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, 13(4):403–437, 2008.

[66] S. T. Gries. Dispersions and adjusted frequencies in corpora: further explorations. *Corpus linguistic applications: current studies, new directions*, pages 197–212, 2010.

[67] S. T. Gries. Corpus data in usage-based linguistics what's the right degree of granularity for the analysis. *Cognitive linguistics: Convergence and expansion*, 32:237, 2011.

[68] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[69] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[70] C. Haney, C. Banks, and P. Zimbardo. Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, 1:69–97, 1973.

[71] A. Hardie. CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3):380–409, 2012.

[72] C. Hemming and M. Lassi. Copyright and the web as corpus. *Retrieved October*, 27:2010, 2010.

[73] C. D. Herrera. Two arguments for 'covert methods' in social research. *The British Journal of Sociology*, 50(2):331–343, 1999.

[74] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. In *UbiComp 2006: Ubiquitous Computing*, pages 177–193. Springer, 2006.

[75] M. Hoey. *Lexical priming*. Wiley Online Library, 2005.

[76] M. Hough, P. Mayhew, and G. Britain. *The British crime survey: first report*. HM Stationery Office, 1983.

[77] M. Hundt, A. Sand, and R. Siemund. *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Albert-Ludwigs-Universität Freiburg, 1998.

[78] M. Hundt, A. Sand, and P. Skandera. *Manual of Information to Accompany the Freiburg-Brown Corpus of American English ('Frown')*. Albert-Ludwigs-Universität Freiburg, 1999.

[79] D. Huynh, D. R. Karger, and D. Quan. Haystack: A platform for creating, organizing and visualizing information using RDF. In *Semantic Web Workshop*, 2002.

[80] N. M. Ide and C. M. Sperberg-McQueen. The TEI: History, goals, and future. In *Text Encoding Initiative*, pages 5–15. Springer, 1995.

[81] Justin.tv. Inc. Watch live video. http://web.archive.org/web/*/http://www.jennicam.org/. Accessed 2014-05-16.

[82] Ustream. Inc. Ustream — the leading HD streaming video platform. http://www.ustream.tv/. Accessed 2014-05-16.

[83] P. G. Ipeirotis. Analyzing the Amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[84] A. N. Jain and A. E. Cleves. Does your model weigh the same as a duck? *Journal of computer-aided molecular design*, 26(1):57–67, 2012.

[85] T. Järvinen. Annotating 200 million words: the bank of English project. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, COLING '94, pages 565–568, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[86] S. Johansson. The LOB corpus of British English texts: presentation and comments. *ALLC journal*, 1(1):25–36, 1980.

[87] S. Johansson. The tagged LOB corpus: User's manual. 1986.

[88] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[89] A. Joulain-Jay. Coping with errors: Estimating the impact of OCR errors on corpus linguistic analysis of historical newspapers. In *ICAME36: Words, Words, Words–Corpora and Lexis*, may 2015.

[90] F. W. Kaeding. *Häufigkeitswörterbuch der deutschen Sprache*. Selbstverlag des Herausgebers, 1897.

[91] G. Kalton. *Introduction to survey sampling*, volume 35. Sage, 1983.

[92] A. Kehoe. Diachronic linguistic analysis on the web with WebCorp. *Language and Computers*, 55(1):297–307, 2006.

[93] A. Kilgarriff. Linguistic search engine. In *proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, pages 53–58, 2003.

[94] A. Kilgarriff. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276, 2005.

[95] A. Kilgarriff. Googleology is bad science. *Computational linguistics*, 33(1):147–151, 2007.

[96] A. Kilgarriff and G. Grefenstette. Web as corpus. In *Proceedings of Corpus Linguistics 2001*, pages 342–344. Corpus Linguistics. Readings in a Widening Discipline, 2001.

[97] A. Kilgarriff and G. Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.

[98] A. Kilgarriff, S. Reddy, J. Pomikálek, and A. Pvs. A corpus factory for many languages. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC-2011), Valletta, Malta*. Citeseer, 2010.

[99] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin. Scholarly context not found: One in five articles suffers from reference rot. *PLoS ONE*, 9(12):e115253, 12 2014.

[100] W. Koehler. Web page change and persistence — a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171, 2002.

[101] W. Koehler. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2), 2004.

[102] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

[103] C. Kohlschtter. Boilerplate: Removal and fulltext extraction from HTML pages. *URL: http://code. google. com/p/boilerpipe*.

[104] K. Kucera. The czech national corpus: principles, design, and results. *Literary and linguistic computing*, 17(2):245–257, 2002.

[105] T. S. Kuhn. The structure of scientific revolutions. *Chicago and London*, 1970.

[106] K. Kwok. Project naptha. http://projectnaptha.com/. Accessed 2014-05-16.

[107] M. Kytö. *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. University of Helsinki, 1993.

[108] M. Källström. Narrative clip – a wearable, automatic lifelogging camera. http://getnarrative.com/. Accessed 2014-05-16.

[109] C. Learning. The Collins corpus. http://www.collins.co.uk/page/The+Collins+Corpus, jul 2015.

[110] D. Y. Lee. Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle. 2001.

[111] D. Y. Lee. The BNC index. http://www.uow.edu.au/~dlee/home/corpus_resources.htm, 2003. Excel spreadsheet.

[112] K. Lee and D. P. Ellis. Voice activity detection in personal audio recordings using autocorrelogram compensation. In *INTERSPEECH 2006: ICSLP: Proceedings of the Ninth International Conference on Spoken Language Processing: September 17-21, 2006, Pittsburgh, Pennsylvania, USA*, pages 1970–1973. ISCA, 2006.

[113] G. Leech. Corpora and theories of linguistic performance. *Directions in corpus linguistics*, pages 105–122, 1992.

[114] G. Leech. 100 million words of English. *English Today*, 9(01):9–15, 1993.

[115] G. Leech. New resources, or just better old ones? The holy grail of representativeness. *Language and Computers*, 59(1):133–149, 2006.

[116] M. D. Lieberman and W. A. Cunningham. Type I and type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, 4(4):423–428, 2009.

[117] J. Lijffijt, T. Nevalainen, T. Saily, P. Papapetrou, K. Puolamaki, and H. Mannila. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 2014.

[118] S. Lohr. *Sampling: design and analysis*. Cengage Learning, 2009.

[119] W. Lund and E. Ringger. Error correction with in-domain training across multiple OCR system outputs. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 658–662, Sept 2011.

[120] M. Madden and A. Smith. Reputation management and social media. 2010.

[121] C. Mair. Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. *Language and Computers*, 55(1):355–376, 2006-01-01T00:00:00.

[122] R. Malaga. Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Research*, 1:403–417, 2001.

[123] S. Mann. Wearable, tetherless, computer-mediated reality (with possible future applications to the disabled)'technical reporti# 260, 1994.

[124] A. McCallum, K. Nigam, et al. A comparison of event models for naïve Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[125] A. McEnery and R. Xiao. Parallel and comparable corpora: What is happening. *Incorporating Corpora. The Linguist and the Translator*, pages 18–31, 2007.

[126] T. McEnery and A. Hardie. Research ethics in corpus linguistics. In *CL2011 Abstracts*, jul 2011. Conference Presentation.

[127] T. McEnery and A. Wilson. *Corpus linguistics: An introduction*. Edinburgh University Press, 2001.

[128] T. McEnery and Z. Xiao. The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Religion*, 17:3–4, 2004.

[129] K. McGrath and J. Waterton. British social attitudes 1983-1986 panel survey, 1986.

[130] F. Meunier, S. De Cock, G. Gilquin, and M. Paquot. *A Taste for Corpora: In honour of Sylviane Granger*. Studies in Corpus Linguistics. John Benjamins Publishing Company, 2011.

[131] S. Meyer zu Eissen and B. Stein. Genre classification of web pages. In S. Biundo, T. Frühwirth, and G. Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3238 of *Lecture Notes in Computer Science*, pages 256–269. Springer Berlin Heidelberg, 2004.

[132] H. Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

[133] R. M. Neal. Slice sampling. *Ann. Statist.*, 31(3):705–767, 06 2003.

[134] M. Nelson and B. Allen. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1):1082–9873, 2002.

[135] J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):pp. 558–625, 1934.

[136] M. Paquot and Y. Bestgen. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers*, 68(1):247–269, 2009.

[137] T. Pedersen. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470, 2008.

[138] J. Pentheroudakis, D. Bradlee, and S. Knoll. Tokenizer for a natural language processing system, Aug. 15 2006. US Patent 7,092,871.

[139] J. Pomikálek. justext: Heuristic based boilerplate removal tool. http://code.google.com/p/justext. Accessed September 2014.

[140] C. U. Press. Cambridge announce new spoken corpus. http://www.cambridge.org/about-us/news/spoken-bnc2014-project-announcement/, jul 2014.

[141] W. T. Preyer. *The Mind of the Child: The development of the intellect*, volume 2. Appleton, 1889.

[142] A. Przepiórkowski, R. L. Górski, B. Lewandowska-Tomaszyk, and M. Lazinski. Towards the national corpus of Polish. In *LREC*, 2008.

[143] M. Rabin et al. *Inference by believers in the law of small numbers*. Institute of Business and Economic Research, 2000.

[144] B. Rampton, J. Channell, P. Rea-Dickens, C. Roberts, and J. Swann. BAAL recommendations on good practice in applied linguistics, 1994.

[145] R. Rapp. Using word familiarities and word associations to measure corpus representativeness. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[146] P. Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008.

[147] P. Rayson. Wmatrix: a web-based corpus processing environment. 2008.

[148] A. Renouf. Webcorp: providing a renewable data source for corpus linguists. *Language and Computers*, 48(1):39–58, 2003.

[149] P. Resnik and N. A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

[150] P. Rossi, J. Wright, and A. Anderson. *Handbook of Survey Research*. Quantitative Studies in Social Relations. Elsevier Science, 2013.

[151] D. Roy. The birth of a word. http://www.ted.com/talks/deb_roy_the_birth_of_a_word. Accessed 2014-05-16.

[152] R. Sanderson, M. Phillips, and H. Van de Sompel. Analyzing the persistence of referenced web resources with Memento. *Arxiv preprint arXiv:1105.3459*, 2011.

[153] M. Santini, A. Mehler, and S. Sharoff. Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York, 2010.

[154] R. Schäfer, A. Barbaresi, and F. Bildhauer. Focused web corpus crawling. *EACL 2014*, page 9, 2014.

[155] R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In *LREC Proc.*, volume 8, 2012.

[156] R. Schäfer and F. Bildhauer. Web corpus construction. *Synthesis Lectures on Human Language Technologies*, 6(4):1–145, 2013.

[157] C. Shaoul and W. C. A reduced redundancy usenet corpus. 2013.

[158] S. Sharoff. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98, 2006.

[159] S. Sharoff. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462, 2006.

[160] S. Sharoff. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*, pages 83–94, 2007.

[161] S. Sharoff. Approaching genre classification via syndromes. In F. Formato and A. Hardie, editors, *Corpus Linguistics 2015 Abstract Book*, pages 306–308, jul 2015.

[162] S. V. Shastri. The Kolhapur corpus of Indian English and work done on its basis so far. *ICAME journal*, 12:15–26, 1988.

[163] J. Sinclair. Reflection on computer corpora in English language research. *Computer Corpora in English Language Research*, pages 1–6, 1982.

[164] J. Sinclair. *Looking Up: An account of the COBUILD Project in lexical computing*. Collins London, 1987.

[165] J. Sinclair. *Corpus, concordance, collocation*. Oxford University Press, 1991.

[166] J. Sinclair. Preliminary recommendations on corpus typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), 1996.

[167] P. Sloan. How to scale Mt. Google. *Business 2.0*, pages 54–55, may 2007.

[168] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):pp. 3–23, 1993.

[169] S. Sridharan and B. Murphy. Modeling word meaning: Distributional semantics and the corpus quality-quantity trade-off. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, COLING*, pages 53–68, 2012.

[170] W. Stern. Psychology of early childhood up to the sixth year. *Trans. by A. Barwell from the 3rd edition) New York: Holt*, page 557, 1924.

[171] M. Stubbs. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1):23–55, 1995-01-01T00:00:00.

[172] D. Summers. Longman/Lancaster English language corpus–criteria and design. *International Journal of Lexicography*, 6(3):181–208, 1993.

[173] M. Taylor, Economic, and C. U. K. R. C. o. M.-S. C. Social Research Council (ESRC). *British Household Panel Survey User Manual: Volumes B1-B5-Codebooks*. 1996.

[174] C. Teddlie and F. Yu. Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research*, 1(1):77–100, 2007.

[175] J. Trost. Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology*, 9(1):54–57, 1986.

[176] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[177] D. C. Tyler and B. McNeil. Librarians and link rot: a comparative analysis with some methodological considerations. *portal: Libraries and the Academy*, 3(4):615–632, 2003.

[178] UK Intellectual Property Office. Exceptions to copyright: Research, oct 2014.

[179] T. Váradi. Corpus linguistics-linguistics or language engineering?'. In *Information Society Multi-Conference Proceedings Language Technologies*, pages 1–5, 2000.

[180] T. Váradi. The linguistic relevance of corpus linguistics. In *Proceedings of the Corpus Linguistics 2001 Conference. UCREL Technical Papers*, volume 13, pages 587–593, 2001.

[181] S. Wallis. That vexed problem of choice. *London: Survey of English Usage, UCL. Retrieved*, 14, 2013.

[182] M. P. West. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green, 1953.

[183] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[184] D. Wu and P. Fung. Improving chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLC '94, pages 180–181, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[185] M. Wynne. *Developing linguistic corpora: a guide to good practice*. Oxbow Books, 2005.

[186] Z. Xiao. Well-known and influential corpora. *A. Lüdeling & M. Kyto (eds) Corpus Linguistics: An International Handbook*, 1:383–457, 2008.

[187] J. Yuan, M. Liberman, and C. Cieri. Towards an integrated understanding of speaking rate in conversation. In *INTERSPEECH*, 2006.

[188] G. K. Zipf. Human behavior and the principle of least effort. 1949.

Appendices

A LWAC Data Fields

```
Data{
  server            : {
    links             : [1, 2, 3]
    complete_sampl...: 2
    complete_samples  : [0, 1]
    next_sample_date  : 1366381980
    current_sample_id: 1
    config            : {
      storage           : {
        root              : corpus
        state_file        : state
        sample_subdir     : samples
        sample_filename   : sample
        files_per_dir     : 1000
        database          : {
          filename          : corpus/links.db
          table             : links
          transaction_limit: 100
          pragma            : {
            locking_mode      : EXCLUSIVE
            cache_size        : 20000
            synchronous       : 0
            temp_store        : 2
          }
          fields            : {
            id                : id
            uri               : uri
          }
        }
      }
      sampling_policy   : {
        sample_limit      : 2
        sample_time       : 60
        sample_alignment  : 0
      }
      client_policy     : {
        dry_run           : false
        fix_encoding      : true
        target_encoding   : UTF-8
        encoding_options  : {
          invalid           : replace
          undef             : replace
          universal_newline: true
        }
        max_body_size     : 20971520
        mimes             : {
          policy            : whitelist
```

```
48            ignore_case       : true
49            list              : ["\^text\\/?.*\$"]
50          }
51          curl_workers      : {
52            max_redirects     : 5
53            useragent         : "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US
                  ...
54            enable_cookies    : true
55            verbose           : false
56            follow_location   : true
57            timeout           : 60
58            connect_timeout   : 10
59            dns_cache_timeout: 10
60            ftp_response_t...: 10
61          }
62        }
63        client_management: {
64          time_per_link     : 5
65          empty_client_b...: 60
66          delay_overesti...: 10
67        }
68        server            : {
69          interfaces        : [{:interface=>"localhost", :port=>27400}]
70          service_name      : downloader
71        }
72        logging           : {
73          progname          : Server
74          logs              : {
75            default           : {
76              dev               : STDOUT
77              level             : info
78            }
79            file_log          : {
80              dev               : logs/server.log
81              level             : info
82            }
83          }
84        }
85      }
86      version           : 0.2.0b
87    }
88    sample            : {
89      id                : 1
90      start_time        : 2013-04-19 15:33:10 +0100
91      end_time          : 2013-04-19 15:33:11 +0100
92      complete          : true
93      open              : false
94      size              : 3
95      duration          : 1.406624844
96      start_time_s      : 1366381990
97      end_time_s        : 1366381991
98      size_on_disk      : 214259.0
99      last_contiguou...: 3
100     dir               : corpus/samples/1
101     path              : corpus/samples/1/sample
102   }
103   datapoint         : {
104     id                : 3
105     uri               : http://google.co.uk
106     dir               : corpus/samples/1/0
107     path              : corpus/samples/1/0/3
108     client_id         : LOCAL3_7ba2f8cd03d79efbbaa4b1c561759c6e
109     error             :
```

```
110       headers            : {
111         Location         : http://www.google.co.uk/
112         Content_Type     : text/html; charset=UTF-8
113         Date             : Fri, 19 Apr 2013 14:33:10 GMT
114         Expires          : -1
115         Cache_Control    : private, max-age=0
116         Server           : gws
117         Content_Length   : 221
118         X_XSS_Protection : 1; mode=block
119         X_Frame_Options  : SAMEORIGIN
120         Set_Cookie       : NID=67=B7dOglOF9YR3BvNje7Xgy_FAHcHIgJMW3HGm9HYI...
121         P3P              : CP="This is not a P3P policy! See http://www.go...
122         Transfer_Encoding: chunked
123       }
124       head               : HTTP/1.1 301 Moved Permanently\nLocation: http:/...
125       body               : <!doctype html><html itemscope="itemscope" item...
126       response           : {
127         round_trip_time  : 0.313531
128         redirect_time    : 0.219863
129         dns_lookup_time  : 0.00129
130         effective_uri    : http://www.google.co.uk/
131         code             : 200
132         download_speed   : 163125.0
133         downloaded_bytes : 51145.0
134         encoding         : text/html; charset=UTF-8
135         truncated        : false
136         mime_allowed     : true
137         dry_run          : false
138       }
139     }
140   }
```

Listing 1: Data stored on each datapoint within an LWAC corpus

B Corpus Postprocessing Scripts

The post-processing tools presented here are largely tasked with transforming the disparate formats recorded by different recording methods into a standard CSV format, based on Lee's BNC index. Where appropriate, fields such as the medium and word count were computed on-the-fly.

In addition to these methods, annotation was aided by manually creating lookup tables to impute metadata for some fields: for example, musical genre is easily and reliably determined by artist.

Files

Files not sourced from a version control system were stored each day. The processing script for these uses the diff tool to compute patches between each version, outputting the word counts to a CSV file broken down per-day.

Phone Calls

Calls were recorded in CSV format by a smartphone application. This script simply reformatted the CSV to be compatible with other records. Due to the low number, calls were tagged by hand using records from the two notebooks.

Emails

Emails were processed from mbox format by normalising their character sets and parsing of their headers to extract timestamps, addressees, and subjects. Email contents were then exported to an intermediate CSV format for further processing (such as removal of replies and word counting), which were then transformed into the same format as other data.

IRC Logs

The bot that monitored IRC rooms already contained a timeout, and would stop listening ten minutes after the last activity by the user of interest. The first 20 characters of the first message beginning a conversation was taken as a reminder of the topic, and word counts were exported to the unified CSV format.

Music (`last.fm`)

A third-party script was used to scrape the entirety of the subject's music-listening history from the 'scrobbling' service `last.fm`. This was returned in TSV format, containing the time at which each track was listened to, along with identifying information. This was converted into the standard CSV format. Later stages computed word counts by automatically searching for song lyrics using a web-based API.

SMS

SMS records were exported in a similar manner to phone calls, and simply converted into the normalised CSV format.

Web logs (SQUID)

SQUID logs were filtered to identify documents which would have been rendered by the subject's browser. This was done by filtering on:

- Username (the subject used a different username for each device);
- MIME type (in order to remove images, CSS, javascript and other non-readable formats);
- URL patterns (in order to remove advertising, AJAX calls, and analytics);
- Return code.

This yielded a list of URLs accessed by each device for each day. These were then retrieved and processed further using Nokogiri to remove boilerplate and normalise character set in order to achieve a word count. Results from this stage were then transformed into the standard CSV format. An additional parameter, the time between page reloads, was also extracted but proved not to be useful due to the low number of reloaded pages.

Terminals

The `script` tool was used to log terminal output. This logs every single instruction provided to the terminal, and as such the main post-processing task was to resolve the meaning of control characters by passing this into a terminal emulator. In order to estimate the proportion of fast-moving output read by the subject, timestamps were used to gauge when terminal output had stopped for more than 60 seconds, above which the last 40 lines of output were taken as a text. This scrollback was chosen based upon the geometry of the subject's screen and constituted a rough average of window size. Word counts from these were then computed for insertion into the normalised CSV format.

C Lee's BNC Genres

C.1 Spoken Genres

Genre	Words	%age	Big Genre	Files
S_brdcast_discussn	757317	7.3%	Broadcast (10.2%)	53
S_brdcast_documentary	41540	0.4%		10
S_brdcast_news	261278	2.5%		12
S_classroom	429970	4.2%		58
S_consult	138011	1.3%		128
S_conv	4206058	40.7%		153
S_courtroom	127474	1.2%		13
S_demonstratn	31772	0.3%		6
S_interview	123816	1.2%	Interviews (9.1%)	13
S_interview_oral_history	815540	7.9%		119
S_lect_commerce	15105	0.1%	Lectures (2.9%)	3
S_lect_humanities_arts	50827	0.5%		4
S_lect_nat_science	22681	0.2%		4
S_lect_polit_law_edu	50881	0.5%		7
S_lect_soc_science	159880	1.5%		13
S_meeting	1377520	13.3%		132
S_parliament	96239	0.9%		6
S_pub_debate	283507	2.7%		16
S_sermon	82287	0.8%		16
S_speech_scripted	200234	1.9%	Speeches (6.4%)	26
S_speech_unscripted	464937	4.5%		51
S_sportslive	33320	0.3%		4
S_tutorial	143199	1.4%		18
S_unclassified	421554	4.1%		44
TOTAL	10334947	100.0%		909

C.2 Written Genres

Genre	Words	%age	Big Genre	Files
W_ac_humanities_arts	3321867	3.8%		87
W_ac_medicine	1421933	1.6%		24
W_ac_nat_science	1111840	1.3%	Academic	43
W_ac_polit_law_edu	4640346	5.3%	Prose (17.7%)	186
W_ac_soc_science	4247592	4.9%		138
W_ac_tech_engin	686004	0.8%		23
W_admin	219946	0.3%		12
W_advert	558133	0.6%		60
W_biography	3528564	4.0%		100
W_commerce	3759366	4.3%		112
W_email	213045	0.2%		7
W_essay_sch	146530	0.2%	Non-printed	7
W_essay_univ	65388	0.1%	Essays (0.3%)	4
W_fict_drama	45757	0.1%		2
W_fict_poetry	222451	0.3%	Fiction	30
W_fict_prose	15926677	18.2%	(18.6%)	432
W_hansard	1156171	1.3%		4
W_institut_doc	546261	0.6%		43
W_instructional	436892	0.5%		15
W_letters_personal	52480	0.1%	Letters (0.2%)	6
W_letters_prof	66031	0.1%		11
W_misc	9140957	10.5%		500
W_news_script	1292156	1.5%		32
W_newsp_brdsht_nat_arts	351811	0.4%		51
W_ newsp_brdsht_nat _commerce	424895	0.5%		44
W_ newsp_brdsht_nat _editorial	101742	0.1%	Broadsheet	12
W_ newsp_brdsht_nat _misc	1032943	1.2%	National	95
W_ newsp_brdsht_nat _reportage	663355	0.8%	Newspapers	49
W_ newsp_brdsht_nat _science	65293	0.1%	(3.5%)	29
W_ newsp_brdsht_nat _social	81895	0.1%		36
W_ newsp_brdsht_nat _sports	297737	0.3%		24
W_newsp_other_arts	239258	0.3%		15
W_newsp_other_commerce	415396	0.5%		17
W_newsp_other_report	2717444	3.1%	Regional	39
W_newsp_other_science	54829	0.1%	Newspapers	23
W_newsp_other_social	1143024	1.3%	(6.4%)	37
W_newsp_other_sports	1027843	1.2%		9
W_newsp_tabloid	728413	0.8%	Tabloids (0.8%)	6
W_non_ac_ humanities_arts	3751865	4.3%		111
W_non_ac_medicine	498679	0.6%		17
W_non_ac_nat_science	2508256	2.9%	Non-	62
W_non_ac_polit_law_edu	4477831	5.1%	academic	93
W_non_ac_soc_science	4187649	4.8%	Prose (19.1%)	128
W_non_ac_tech_engin	1209796	1.4%		123
W_pop_lore	7376391	8.5%		211
W_religion	1121632	1.3%		35
TOTAL	87284364	100.0%		3144

D BNC68 Model Accuracy

```
Correctly Classified Instances        2426               71.4791 %
Incorrectly Classified Instances       968               28.5209 %
Kappa statistic                          0.7029
Mean absolute error                      0.0084
Root mean squared error                  0.0912
Relative absolute error                 29.6594 %
Root relative squared error             76.7327 %
Total Number of Instances             3394

=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
0.352     0.021      0.384     0.352     0.368      0.708    W_non_ac_soc_science
0.917     0.001      0.846     0.917     0.88       0.956    W_ac_medicine
0.8       0.004      0.5       0.8       0.615      0.926    W_instructional
0.706     0.001      0.75      0.706     0.727      0.988    W_newsp_other_commerce
0.76      0.001      0.864     0.76      0.809      0.953    S_speech_scripted
0         0          0         0         0          0.459    W_fict_drama
0.875     0          0.933     0.875     0.903      0.964    S_pub_debate
0         0          0         0         0          0.895    S_brdcast_documentary
1         0          1         1         1          1        W_email
0.811     0.009      0.793     0.811     0.801      0.982    S_meeting
0.73      0.011      0.676     0.73      0.702      0.879    W_biography
0.606     0.017      0.614     0.606     0.61       0.854    W_ac_soc_science
0         0          0         0         0          0.79     S_demonstratn
0.545     0.001      0.667     0.545     0.6        0.949    W_letters_prof
1         0.005      0.6       1         0.75       0.998    W_newsp_brdsht_nat_sports
0.744     0.003      0.762     0.744     0.753      0.887    W_ac_nat_science
0.5       0.001      0.75      0.5       0.6        0.829    S_tutorial
0.425     0.013      0.463     0.425     0.443      0.741    W_ac_humanities_arts
0.523     0.019      0.487     0.523     0.504      0.777    W_non_ac_humanities_arts
0.774     0.009      0.586     0.774     0.667      0.968    S_brdcast_discussn
0.783     0.003      0.621     0.783     0.692      0.926    W_ac_tech_engin
0.922     0.009      0.618     0.922     0.74       0.995    W_newsp_brdsht_nat_arts
0.917     0.001      0.786     0.917     0.846      0.999    W_newsp_brdsht_nat_editorial
0.821     0.007      0.561     0.821     0.667      0.935    W_newsp_other_report
0.714     0.008      0.472     0.714     0.568      0.881    W_religion
0.111     0          1         0.111     0.2        0.982    W_newsp_brdsht_nat_social
0.429     0.001      0.6       0.429     0.5        0.83     S_lect_polit_law_edu
0.714     0          0.833     0.714     0.769      0.841    W_essay_school
0.733     0.003      0.5       0.733     0.595      0.957    W_newsp_other_arts
0.969     0.001      0.912     0.969     0.939      0.997    W_news_script
0         0          0         0         0          0.461    W_essay_univ
0.627     0.013      0.427     0.627     0.508      0.934    S_speech_unscripted
0.417     0.003      0.357     0.417     0.385      0.817    W_admin
0.457     0.008      0.623     0.457     0.528      0.736    W_non_ac_polit_law_edu
0.667     0.001      0.571     0.667     0.615      0.813    W_newsp_tabloid
0.833     0          1         0.833     0.909      0.999    W_letters_personal
1         0          0.8       1         0.889      1        W_hansard
0.765     0.004      0.481     0.765     0.591      0.928    W_non_ac_medicine
0.909     0.001      0.909     0.909     0.909      0.999    W_newsp_brdsht_nat_commerce
0.972     0.008      0.946     0.972     0.959      0.986    W_fict_prose
0.651     0.009      0.475     0.651     0.549      0.911    W_institut_doc
0.211     0.005      0.571     0.211     0.308      0.927    W_newsp_brdsht_nat_misc
0.705     0.014      0.763     0.705     0.733      0.876    W_pop_lore
0.276     0.001      0.667     0.276     0.39       0.989    W_newsp_brdsht_nat_science
0.891     0.003      0.906     0.891     0.898      0.991    S_interview_oral_history
0.769     0.002      0.625     0.769     0.69       0.949    S_lect_soc_science
0.833     0          0.833     0.833     0.833      0.908    S_parliament
0.889     0.004      0.348     0.889     0.5        0.936    W_newsp_other_sports
0.615     0          1         0.615     0.762      0.894    S_interview
0.25      0.001      0.333     0.25      0.286      0.846    S_sportslive
```

0.914	0.002	0.936	0.914	0.925	0.996	S_consult
0.9	0.002	0.771	0.9	0.831	0.997	W_fict_poetry
0.324	0.006	0.364	0.324	0.343	0.842	W_newsp_other_social
1	0.001	0.889	1	0.941	1	S_sermon
1	0	0.929	1	0.963	1	S_courtroom
0.667	0.001	0.5	0.667	0.571	0.823	S_lect_commerce
0.136	0.003	0.353	0.136	0.197	0.862	S_unclassified
0.959	0.001	0.967	0.959	0.963	0.978	W_non_ac_tech_engin
0.98	0.013	0.522	0.98	0.681	0.994	W_newsp_brdsht_nat_report
0.661	0.002	0.848	0.661	0.743	0.975	W_advert
0.75	0.003	0.474	0.75	0.581	0.939	S_brdcast_news
0.25	0	0.5	0.25	0.333	0.702	S_lect_humanities_arts
0.5	0	1	0.5	0.667	0.85	S_lect_nat_science
0.705	0.008	0.76	0.705	0.731	0.883	W_commerce
0.348	0	1	0.348	0.516	0.988	W_newsp_other_science
0.887	0.013	0.567	0.887	0.692	0.94	W_non_ac_nat_science
0.71	0.009	0.825	0.71	0.763	0.861	W_ac_polit_law_edu
0.793	0.003	0.821	0.793	0.807	0.973	S_classroom
0.715	0.008	0.724	0.715	0.704	0.913	(Weighted Avg.)

E BNC45 Model Evaluation

Though trained on BNC data, the majority of data classified here is written directly for the web. The impact this has upon classifier performance is not known, and insufficient gold standard data exists, tagged at sufficient resolution, to perform a direct evaluation.

Instead, data indexed by the Open Directory Project (DMOZ) was used as a measure of systematic bias. The classification scheme used in DMOZ is significantly more diverse than that used in Lee's BNC index, covering a wider range of topics to a finer level of granularity. As its goal is to represent a greater portion of the population than simply British English users this is unsurprising, however, it is organised hierarchically, and thus has larger-scale categories which are more compatible.

For each of the documents retrieved in the final corpus, an attempt was made to identify its DMOZ-equivalent category. This was performed by matching URL or, failing such a detailed pattern, domain name, against 699,385 URLs taken from the publicly-accessible DMOZ archive.

For each BNC category (as assigned by the BNC45 model used in the retrieval task), the homogeniety of DMOZ categories was measured at each level of the hierarchy. Categories below those which consist only of a single uppercase letter were concatenated, in order to remove some 'A-Z' listing categories.

The full set of 55,790 files downloaded for the BNC corpus comparison in Chapter 6 were used as source data. Of these, 777 matched full-URL categories from the DMOZ classification, with a further 32,697 matching domain only.

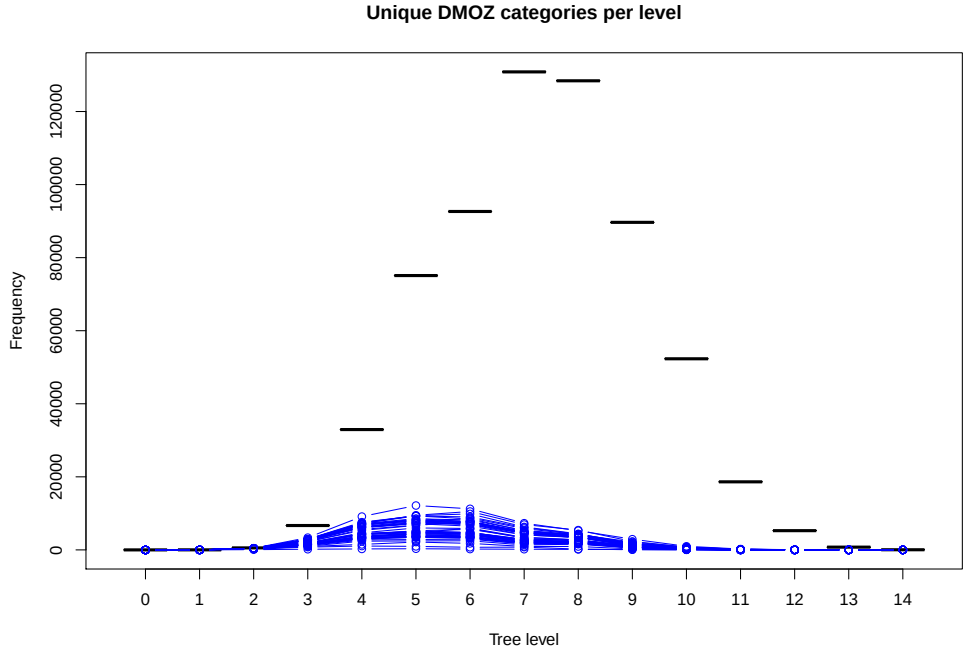


Figure 1: Unique category counts within and between BNC genres.

The number of unique categories in the DMOZ overall is shown in Figure 1. The black lines illustrate the number of unique DMOZ categories at each level of the tree; the blue lines show the number of unique categories within each BNC genre, as classified by the BNC45 classifier model.

Figure 2 shows log transformed variance in a similar manner. Under ideal circumstances, where the BNC45 classifier is in perfect agreement with one level of the DMOZ across all categories, the variance remaining within each classifier will drop to zero at a given tree level. Though this occurs for some values at a tree depth of around 10 for most genres, this is well beyond the bell curve of DMOZ classification, indicating that fewer DMOZ categories descend beyond this depth anyway.

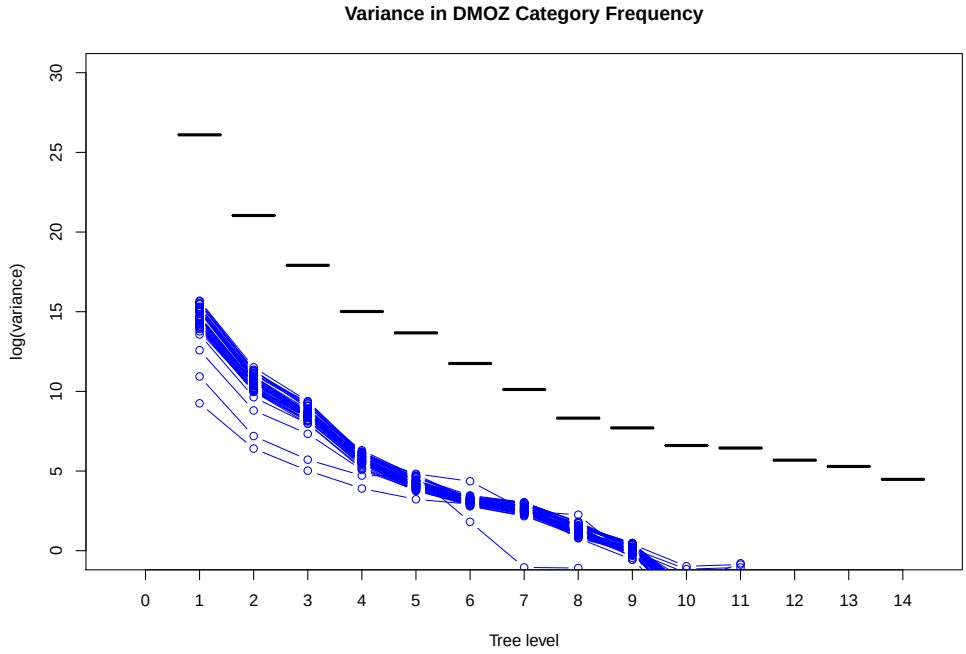


Figure 2: Log variance in DMOZ category frequency across and within BNC genres

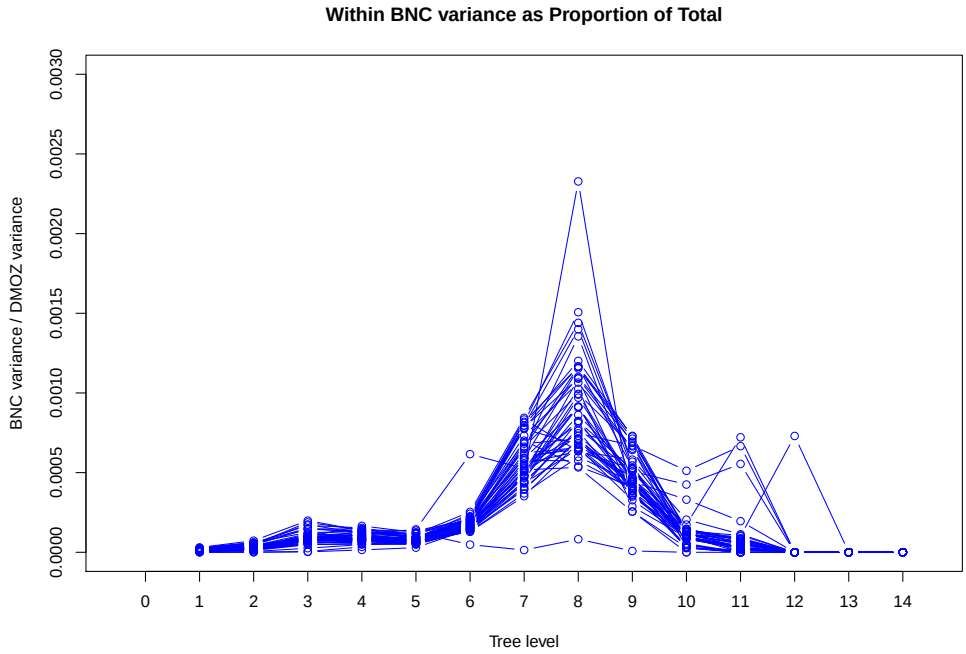


Figure 3: Within-genre DMOZ category frequency variance as a proportion of global variance

Figure 3 shows the proportion of variance within each BNC genre. This displays a small reduction around a tree depth of five, mirroring the steepening of the curve in Figure 2. This eclipsed by the relatively poor fit of later categories (as would be expected), before dropping off due to the smaller frequency of DMOZ categories beyond a tree depth of approximately ten.

These plots provide tentative evidence that best agreement between the DMOZ classification and the BNC45 classifier exists at a tree depth of around five. It also indicates that the DMOZ classifications are potentially a useful tool in guiding retrieval of data, or, given sufficient human alignment, construction of gold standard data for further classifier evaluation.

F Retrieved Corpus Keywords

F.1 Under-represented Keywords

(Continues on next page)

46513	she'd	7338	4050	17407.2730038145
32039	darlington	5559	1521	16656.975047457
20600	ec	6091	5531	11220.7494337932
7882	solicitors	2135	1086	5236.9941851402
35028	swindon	2182	1409	4825.7478738986
11991	kinnock	1375	264	4466.7414274941
99221	theda	838	32	3297.2583100254
289711	nizan	756	6	3146.0766292512
9792	cnaa	742	1	3140.1387559493
89224	guil	755	43	2886.4701888914
30750	maastricht	1165	593	2856.8794199197
67028	oesophageal	957	279	2820.0060616266
28560	lexical	1106	514	2809.1229318242
121272	robyn	1200	707	2767.4801890584
16530	lamont	1184	708	2712.6352767427
92552	risc	1013	416	2689.3520456627
10850	knitting	1509	1463	2668.8854589916
208041	athelstan	1060	516	2645.2960327523
37406	privatisation	1104	665	2521.089781504
339175	leapor	626	18	2502.1514682347
12424	heseltine	803	206	2445.4799531759
154162	pylori	1105	727	2422.2250112204
78345	d'you	704	126	2318.6575844909
100322	sparc	837	330	2252.8796969432
1305	eec	1072	805	2198.891976148
3890	polytechnics	603	59	2181.1750965425
218938	endill	515	1	2175.6275965367
219284	burun	551	27	2131.6415037429
165631	ronni	538	21	2113.9699608352
3068	polytechnic	1070	881	2087.3407463415
86271	morrissey	748	264	2085.9920809487
79902	mips	766	302	2061.7914941895
23181	cranston	769	310	2054.4855423266
58927	syntactic	875	502	2041.5769743531
55871	modigliani	594	91	2012.1916535193
63999	minton	747	310	1976.0245535251
112332	pascoe	672	199	1971.8225088017
88734	shiona	479	9	1949.3640749749
68040	icl	742	329	1917.4414898386
219465	yanto	447	2	1875.5783123003
224397	jaq	449	7	1838.635498068
321530	molla	548	104	1784.342938255
100493	sunsoft	476	34	1782.8397134693
200139	nordern	425	2	1782.2321490091
198080	gedge	469	30	1775.0432169105
120042	leonora	614	201	1751.3847573564
13036	mcallister	723	383	1744.6193691341
112072	souness	510	81	1716.8770520779
35818	lothian	987	982	1716.7954395472

23055	mansell	687	338	1707.404764739
222032	rincewind	408	4	1690.8598482008
223122	penry	480	61	1675.4189061189
171430	ceau	447	32	1673.8527612241
20800	teesside	808	611	1651.2799548777
154167	gastrin	594	222	1626.988736276
120040	fenella	461	55	1624.0391557818
3602	leas	667	355	1606.3860682303
127314	nme	625	288	1592.4733285733
55736	belinda	697	424	1584.6112543979
113025	wilko	399	13	1584.4726764552
100339	osf	730	500	1569.2416891438
230633	colonic	830	724	1566.1873434091
176323	seb	653	353	1562.5809771303
91562	svqs	373	3	1551.8570783571
198167	fabia	448	63	1539.3362497451
99517	microsystems	798	677	1530.3946570334
88661	mungo	573	236	1519.6943478497
64599	fitzalan	497	129	1509.2052824261
41898	ashdown	531	179	1501.5847848396
143128	duodenal	654	397	1488.2766124975
19550	enquired	765	640	1478.9111867089
205786	jessamy	397	30	1478.5548495311
36758	leith	792	704	1477.8719975854
42943	branson	683	467	1469.4588482299
63830	infinitive	695	494	1466.6099058801
214319	defries	357	5	1466.5398249071
171431	escu	417	54	1451.4003095108
56155	alexei	732	589	1446.4547951509
227972	eqn	413	54	1435.3702594557
12604	delors	455	104	1424.0277425677
25028	strathclyde	612	354	1422.8597632183
30718	ici	768	704	1406.6951837265
100494	svr4	408	57	1403.3723114021
89375	belville	367	23	1391.5615431422
162631	vitor	448	108	1384.9508357511
212864	gazzer	358	20	1370.866846458
83921	hcima	338	7	1370.5857061186
30675	gatt	553	276	1366.6957208534
204325	nicolo	418	82	1351.9953348697
64128	oesophagus	482	167	1351.8110548252
89167	insp	517	234	1326.0393963434
22974	sotheby	690	589	1318.5620825021
113081	tbsp	545	287	1318.2932737856
211191	gurder	311	1	1309.193884181
220408	nicandra	311	1	1309.193884181
30966	mitterrand	607	420	1298.2946195841
74666	bernice	668	557	1293.8313649678
11921	dhss	414	96	1291.4604814421

49219	fergie	513	245	1289.6338862787
13069	multiparty	489	211	1276.0217318113
12455	hurd	646	537	1253.412044175
204555	onie	364	52	1247.5268807116
290811	halling	324	19	1235.6557250555
110698	mcleish	384	80	1226.5621745476
12112	gnp	620	502	1220.9023762858
211184	grimma	282	1	1186.0766079215
110489	klerk	501	280	1182.1804988214
211183	nomes	330	36	1177.0476740713
1307	creggan	311	20	1176.4824205129
86136	langbaurgh	286	4	1174.9216059885
88135	flavia	402	125	1163.7265994083
965	wea	459	217	1158.3296903891
27533	althusser	436	180	1155.406530404
13560	devlin	579	461	1150.8067219508
196846	zuwaya	275	2	1146.1439044569
43523	masai	406	138	1145.2658257751
100293	80486	315	30	1143.2007640625
203317	ranulf	402	135	1138.0806329191
100508	usl	511	332	1126.8645036598
12423	reshuffle	548	410	1126.238370085
24106	mellor	584	491	1125.8081025956
34450	microcomputer	438	198	1123.9210848097
36230	yugoslav	606	545	1122.8966251638
204175	lorton	332	52	1120.3759344914
125186	tuppe	264	1	1109.6693542408
82632	ianthe	344	72	1097.7093519398
30710	hercegovina	300	28	1091.4192057629
62217	culley	341	72	1086.0976778548
18310	rostov	444	228	1084.9782914496
30955	redundancies	493	323	1082.8469772842
38606	fundholding	257	1	1079.9578585363
29221	reuter	585	530	1079.1762758644
118499	crilly	284	19	1070.4257931008
46470	abberley	271	10	1068.521720694
24143	computerised	494	335	1067.1344745558
176647	russon	261	5	1061.4388841553
54065	cegb	259	5	1053.0107519231
97772	drily	383	142	1051.7399333255
85747	edouard	505	369	1050.7270744768
174770	huy	404	177	1048.5066722277
79209	bgs	308	46	1047.8682927086
42681	ludens	277	19	1041.5782685994
117068	portadown	362	119	1031.3165314635
37515	arbitrage	418	211	1028.4454918124
325812	cnut	432	238	1025.574997898
30716	tncs	284	29	1021.7800058902
204638	tallis	413	212	1009.3800047482

202535	emmie	285	34	1004.0246917932
16278	wlr	378	156	1001.8271338618
130325	dalgliesh	317	70	999.833889947
30469	cpsu	384	172	988.6703818644
17548	bukharin	312	69	983.7342080563
14910	ilp	401	208	975.9452369735
48098	aycliffe	361	141	974.7024133205
40144	esrc	352	130	967.8095347718
28389	lavatory	486	389	963.1579597726
197755	oesophagitis	269	30	956.615630936
113026	cantona	317	87	949.109509815
71858	gummer	278	42	943.9689603487
100311	cose	358	151	941.6891715476
230487	pepsinogen	241	12	931.2350322132
16694	bureaux	364	167	928.9601467817
142192	tentacle	420	272	927.6028504556
31295	ucta	252	22	923.9788003166
63629	watercolour	405	251	912.9842445472
40753	fmln	271	44	908.7521685175
36079	csce	299	78	906.8142603144
63752	deictic	251	25	905.9048565845
124430	dougal	325	115	905.6316821957
37335	olivetti	350	155	904.8511439067
159601	kaifu	229	10	893.2329107526
85434	feargal	214	2	887.7610373665
3882	privatised	403	270	875.7061082432
214698	aggie	332	141	871.1997114017
255460	coeliac	317	118	869.3918871822
146887	papillae	346	167	866.3030844695
49006	maisie	385	243	860.620143417
193744	titford	209	3	857.9590687749
60370	batty	445	375	856.7158972126
67150	faecal	447	381	854.937945579
222033	twoflower	206	2	853.8954139195
87436	kirov	370	218	853.2919716053
2840	tuc	421	324	852.3602885807
211560	thorfinn	262	50	852.1517211822
40349	banbury	405	290	851.4526208835
91430	nixdorf	242	30	847.8073083532
203970	rohmer	347	180	844.5022929563
34465	ferranti	263	54	842.6589670675
85371	d'arcy	349	185	841.9196603618
293188	rowbottom	233	23	841.8836704678
16735	offeror	361	208	840.7302666109
208341	trunchbull	210	6	839.6446811848
118184	linfield	292	95	834.4097265869
41170	teesdale	291	95	830.7217480438
79657	tranmere	341	177	829.6870793315
12021	conveyancing	298	107	826.6361729786

sorted-keywords

59081	phonological	387	269	825.8683863476
253190	ribber	208	7	824.512667839
31777	dti	415	334	819.9543644193
112460	4oz	239	34	819.676276513

F.2 Over-represented Keywords

(Continues on next page)

sorted-keywords

	key	ref	corp	loglik
4038	p	29217	8839489	1977900.52660189
2642	s	17778	1903037	357245.225266159
2433	she	290860	916731	136365.145796708
67356	gt	111	535127	134335.149383536
22924	com	172	465891	116044.313798295
157415	www	2	448772	113981.587985684
80073	amp	668	432633	102780.694483923
16480	lt	792	425436	99921.7604419157
176	had	359670	1439973	94877.1125714251
7917	h	8094	555374	90772.7204225253
8230	t	9078	557616	87217.0878646232
2449	her	277189	1070672	80897.8098257682
34384	2015	25	293434	74145.0411888397
60062	2014	16	254997	64517.3774052297
35	he	523177	2629585	59109.7836268183
259	was	722425	3866921	58398.9848473108
1610	cent	34353	44410	49457.429430098
13	the	5021789	33346038	47673.9177485614
122	be	524709	2771036	45819.9483379022
119	it	727487	4058706	45306.4107823409
7957	u	2941	234264	40360.2550871008
34722	online	190	148152	35540.504089824
98	to	2117911	13674213	31847.9642452881
7853	m	14132	363108	31792.7042610156
2199	labour	24780	36860	31665.7807527362
1291	n	7761	261109	29021.9005887854
128	but	348119	1867976	27735.7691112495
299	been	215570	1057790	27255.5238390625
10458	posted	607	123889	26385.6717154975
56	would	186520	897188	25753.7355294291
1608	1	77239	978909	24438.8702405685
4503	d	12872	302164	23978.6651970909
89961	org	9	94015	23741.5263324715
104	there	221631	1133637	22956.0637299044
24361	2013	19	91340	22928.7780286913
22278	center	468	104270	22485.1195892446
2406	percent	405	102171	22390.9537410234
244	which	305003	1669688	21504.230066248
9098	ve	376	97206	21366.6173723388
10689	2	944	111027	21327.5105395567
13236	don	1526	120813	20772.3084171785
4712	st	11601	264992	20367.3173630447
63711	okay	882	103794	19940.7618264792
59149	2012	11	78368	19741.7609552155
3211	looked	28796	74087	19281.2997082316
34383	2010	69	78278	19073.9471988043
67120	schmitt	30	74663	18569.1510139359
165501	php	3	72195	18289.9134997198

10562	1	1972	117238	18084.869271963
7183	l	10694	239228	17929.6496922032
1425	said	160817	811561	17816.3916444048
38450	2011	20	71070	17775.8523503162
28274	cc	417	83270	17677.2543087766
970	could	118255	561778	17214.51662021
691	e	19888	338706	16940.0657328758
11586	didn't	28683	79868	16894.6020010591
1609	per	46167	163448	16812.9087221618
8312	san	2118	113626	16731.9238727728
9237	yesterday	17367	34066	16722.9599022924
926	o	5057	155788	16142.2339058857
30214	ll	106	67899	16120.686072945
6482	1989	13001	20490	15749.065910893
60267	2008	28	61093	15155.3558535457
6548	1990	14475	26662	14961.6923175115
604	britain	21299	53550	14772.1968860858
100838	internet	93	62023	14758.9493420349
5272	web	557	74479	14720.5005460966
19552	he'd	9328	10738	14674.3186132746
500	programme	13805	25052	14501.660183452
60273	2009	9	57136	14380.0783642769
2359	behaviour	11168	16520	14342.9884519028
660	de	12927	242351	14234.094036467
60107	2007	20	55224	13759.7057816889
1367	don't	48437	189354	13675.3414229932
3758	0	17536	287984	13465.4378973914
390	they	277024	1600880	13360.2249742561
21397	hon	10573	16156	13183.6322758252
4511	f	7287	168372	13117.8556405327
59395	2006	28	52657	13020.2117033705
100348	pdf	37	52027	12766.2031917794
679	city	19308	301208	12716.2086777529
263	towards	23488	68694	12693.5852502438
2526	thought	42519	163617	12537.0481890215
5339	11	14007	242791	12528.7242375897
5537	him	133314	694782	12520.0006062516
636	were	257618	1488620	12431.8508863345
13281	outlet	475	62718	12362.9761916747
22127	program	3836	118764	12357.1580425946
1734	voice	22261	64254	12317.8303111905
38710	mg	1261	77194	12056.7911606805
3058	2	63505	716529	11830.6515447561
535	then	107971	546791	11756.8535537536
6277	mr	57650	250805	11700.309549642
121	not	349209	2105913	11668.2232539898
33971	henderson	584	61940	11600.2415631191
807	site	7610	162921	11553.6664967085
10	his	370597	2253100	11443.1279539566

12118	winter	5346	133826	11402.2194786503
45693	canberra	135	49830	11368.7506560405
156587	didn	11	45042	11286.2971583163
6564	1991	12138	24532	11283.2139628911
177882	obama	1	44277	11230.9905180448
876	felt	23559	74566	10952.6456937835
12545	la	3933	112066	10847.0098233304
525	up	166593	923692	10814.9990579084
9493	eyes	25703	85465	10784.4115455171
34382	2005	92	45927	10733.194703689
10442	3	457	55234	10677.3043738712
19474	couldn't	11705	24034	10675.3034475423
10412	4	661	58927	10517.5992586374
19676	wasn't	13931	33075	10504.5740980383
3061	3	47731	556230	10427.7818452401
315997	quot	1	40954	10386.8210967182
78778	2004	16	41719	10384.4971469053
1330	your	84463	882537	10244.0991813595
26	social	38710	155064	10195.018962811
79913	email	44	42060	10181.9625267388
10413	7	461	53174	10167.4050217064
170747	ppr	14	40385	10069.679878736
469	if	182163	1033843	10063.1350643265
6379	1988	9975	19015	9924.600514704
1672	seemed	20315	62756	9908.5948487351
1088	council	24729	84884	9672.9216994253
10511	5	1299	66756	9627.4731896512
1587	re	10477	183397	9626.7350517409
38486	gene	2143	79768	9535.0963094793
4925	york	7594	149997	9527.1136248179
2364	down	69869	337787	9437.5290515232
1234	university	12516	204204	9429.2467017681
3776	10	28789	367849	9415.7296093792
826	too	55204	252771	9310.2154072228
10558	6	403	48127	9277.7939492494
4292	c	24941	328772	9223.1099889139
22242	behavior	77	39241	9183.8412990765
49199	smiled	6729	9348	9129.0893556203
4770	12	17211	250936	9127.8801433412
6378	pupils	7208	10927	9053.63488411
3833	ii	8284	154732	9035.601558152
10345	8	549	50188	9024.5414769981
10288	9	661	52420	9018.0418626443
8771	minister	22806	78103	8967.1700836324
59059	2003	58	37636	8942.1587389756
12875	hadn't	6604	9218	8923.3445141712
8689	herself	14960	41473	8877.7035019803
569	should	82486	418714	8877.454754204
957	although	36764	151610	8875.7351446572

1138	back	76751	383996	8865.3435531673
59033	2002	44	36752	8845.1427105891
2024	government	55614	259021	8776.2058424921
1022	only	121269	661568	8738.5798372518
426	rather	35119	143849	8655.9326349766
1347	perhaps	26568	99037	8520.0066543793
457	out	156597	892503	8388.3678176077
8610	esq	83	36179	8370.4439729923
35516	mutations	443	45119	8361.0360806355
4862	defense	182	38921	8345.3349838027
191891	google	2	32506	8225.0168264559
3474	al	6869	132669	8153.6832193832
7161	roll	2574	78930	8146.8559815148
1167	our	67056	700837	8145.8879198004
6148	door	20029	67573	8142.1111855015
42310	doesn	4	32244	8129.7762181112
38451	2001	115	36152	8121.6972912228
53661	color	111	36005	8114.4114237714
22243	labor	172	37642	8097.2629513604
1388	face	30417	121508	8075.4720441463
129	i	453418	3843181	8033.1887731992
87	have	338692	2111469	8001.4133918609
1944	guard	2581	78146	7979.7650943651
2776	party	37214	159656	7916.7781976436
6229	toward	1081	54056	7689.9141870756
3069	5	39444	449040	7687.4323025625
39	that	820530	5452708	7679.7313980869
12977	california	2245	71040	7525.2406735645
1108	no	159426	925114	7442.6277855745
7931	w	5683	114255	7442.6218753895
17710	hills	2349	71704	7372.7568449712
39294	francisco	767	47081	7361.8811444692
3115	canada	2382	71880	7319.0022306031
394	concerned	13320	38612	7314.3035696557
445	into	131868	748012	7312.103479753
419	over	110970	614181	7289.9067057829
112998	nike	62	30939	7230.2878518921
1340	wouldn't	9557	22830	7143.9056159501
6543	1987	8179	17405	7137.0274816702
79894	info	227	34627	7025.5104321713
2046	british	30950	130196	7020.4316304055
1170	centre	18661	65336	6975.5912596027
1638	might	47608	225872	6967.900401166
87801	lauren	116	31339	6924.6266495593
1318	you're	18489	64728	6912.7355965327
1031	asked	26993	109262	6894.2195257067
81650	eu	44	28894	6869.6952758842
7405	award	14084	43540	6859.4835182684
3193	14	11433	172454	6775.470323999

7910	j	9417	151092	6752.4100753252
1919	round	22633	86734	6748.6428618275
1345	13	10837	165186	6639.8471085503
2580	colour	9095	22091	6637.9822289633